MOLECULAR STUDIES OF A DISPERSED, SIMPLE DNA SEQUENCE IN *DROSOPHILA MELANOGASTER*

by

PAUL REDPATH SIMPSON

A thesis submitted for the degree of
Doctor of Philosophy of
The Australian National University.

November, 1984

STATEMENT

The work described in this thesis was performed by myself except where due reference is made in the text. No material in this thesis has been presented for any other degree or diploma.

Paul R. Simpson

ACKNOWLEDGEMENTS

I would like especially to thank my supervisor, Dr George Miklos, for his continued guidance and support over the past three and a half years. His enthusiasm for science and his eagerness to share his knowledge have not failed to stimulate further my own interest in biology.

I am also grateful to Professor Bernard John and Drs John Oakeshott and Robyn Russell for their willingness to read and comment upon all, or parts, of my thesis and to Drs Ian Boussy and Masatoshi Yamamoto and Marion Healy for stimulating discussions.

Bronwyn Matheson, Julie Higginbotham and Valmai Hicks provided excellent technical assistance when requested together with some amazing conversations.

Garry Brown drew most of the illustrations and members of the Photography Unit gave their advice and help to the preparation of many of the figures. David Sandilands and David Smith revealed the potential of the computer facilities for DNA sequence analysis and Erica Batt introduced me painlessly to the intricacies of word processing.

Financial support during the course of this thesis was made possible through an Australian National University Ph.D. Scholarship.

Finally, but most appreciatively, I wish to acknowledge my parents who unwaveringly sent their love and encouragement over a distance of 10,000 miles for a period of 4 years.

ABSTRACT

In order to provide an intensive study of a repeated sequence family, a majority of the dispersed copies of the Garden of Eden, or GOE, family has been isolated from the *Drosophila melanogaster* genome. As well as being present in the genomes of several invertebrate species, GOE sequences are also predominantly associated with the W or Y sex chromosomes of snakes, birds and mammals.

Seven distinct copies from this family have been collected and the sequences of five are analysed here.

All sequences share regions rich in tandemly arranged GATA tetranucleotides. These 'GATA' regions range in length from 90 to 400 nucleotides. As there is no significant homology between their surrounding sequences, the GATA regions alone must correspond to the GOE sequence, or *element*. Analysis of these GOE elements suggests that they are most likely to have arisen from poly(GATA) sequences that accumulated mutations (nucleotide substitutions, deletions or insertions) at random. In addition, a pair of equivalent GOE elements from two *D. melanogaster* strains (Canton S and a wild strain) are identical except for an extra GATA unit in one copy that was probably generated by unequal crossover.

Previous suggestions that GOE element-like sequences are involved in sex determination are discussed in the light of the *Drosophila* data and it is concluded that a direct role is unlikely. Instead, GOE elements seem to belong to a general class of dispersed and simple sequences that are distributed in many eukaryote genomes.

ABBREVIATIONS

bp:	base pair(s)
Ci:	Curie
EDTA:	ethylenediaminetetracetic acid, disodium salt
GOE:	Garden of Eden (a family of repeated sequences)
IPTG:	isopropyl-beta-D-thiogalactopyranoside
kb:	kilobase (pairs)
NaAc:	sodium acetate
pfu:	plaque forming unit
SSC:	standard saline citrate
Tris:	tris(hydroxymethyl)-aminomethane
UV:	ultraviolet
X-gal:	5-bromo-4-chloro-3-indolyl-beta-D-galactopyranoside

TABLE OF CONTENTS

Chapter 1

INTRODUCTION

The presence of repeated DNA sequences has been demonstrated in the genomes of all eukaryotes studied (Schmidtke and Epplen, 1980) and in those of some prokaryotes, in both eubacteria (e.g. Ohtsubo and Ohtsubo, 1977) and in archaebacteria (e.g. Sapienza and Doolittle, 1982). To what extent, then, is the expression of a genome dependent upon, or affected by, these repeated elements?

Conceivably, repeated sequences can act in *cis*, in which case they would need to be both as numerous and as dispersed as genes in order to influence them, or they can act in *trans*, via a soluble intermediate, where these conditions need not apply. Both modes of action require that there be several sets, or families, of related sequences so that different parts of the genome can be coordinately expressed.

The numbers, organisation and variety of repeated sequences can be determined at the gross level by reassociation experiments, although these parameters need not always be related to function. To determine the functional role, if any, of repeated sequence families in the genome, it is necessary to treat them separately, and this is at present only possible through the use of recombinant DNA technology. The following review aims to demonstrate the variety of

organisations and structures that repeated sequences may have and to show why the strategy employed in this thesis is particularly apposite for investigating repeated sequence structure and function.

1.1 The repeated component of eukaryote genomes

Thirty percent of most animal genomes is composed of repeated sequences (Schmidtke and Epplen, 1980), though values as high as 80% for the salamander, *Triturus cristatus* (Baldari & Amaldi, 1977) and as low as 10% for the macronuclear genome of the ciliate, *Tetrahymena pyriformiis* (Borchsenius et al., 1978) have been reported. Only 2% of the genome of the mould, *Aspergillus nidulans* is repeated (Timberlake, 1978), while a higher fraction (usually 70%) of plant genomes is made up of repeated DNA (Flavell et al., 1974).

These values are obtained by measuring the rate at which denatured genomic DNA reassociates to form duplex molecules. Were all sequences to be of one type only (i.e. the genome is composed entirely of unique sequences), the rate of reassociation would be governed by second order kinetics. However, over a given length of time, a denatured strand of a repeated sequence is more likely to find a complement than is that of a unique sequence. Thus, one fraction of the genome reassociates faster than the rest. Theoretically, it is possible to determine the frequency of repetition of a repeated sequence,

that is, the number of copies of the repeated sequence relative to the unique sequences. Practically, because there are many types or families of repeated sequences with differing repetition frequencies, they are often grouped into three or four 'repetition frequency' components (e.g. Britten and Kohne, 1968 and Lewin, 1980). There are no precise limits to these components, though they may be usefully defined in the following way:

The unique component includes those sequences that are present in one or a few copies in each haploid genome.

The highly repeated or satellite component contains those sequences that are present in upwards of a million copies per haploid genome genome.

The middle, moderately or intermediate repeated component incorporates all repetition frequencies that are not included in the two previous components. Consequently, sequences that are present in as few as ten copies (Emmons et al., 1983) and as many as 10^5 copies (Schmid and Jelinek, 1982) can be classed as being 'middle repeated'.

In addition, a small fraction of the genome often re-associates too rapidly to be resolved experimentally. This is due to inverted, homologous sequences that lie on the same DNA strand. These are able to undergo *intra*molecular reassociation, which is more rapid than *inter*molecular reassociation. Such 'foldback' or 'snapback' sequences may be distinct types of repeated sequence, as are the foldback (FB) elements described by Potter (1982), or they may simply be

inverted pairs of homologous repeated sequences that belong to one of the above repetition frequency components.

Highly repeated DNA is commonly associated with satellite sequences (Singer, 1982a). These are characterised by being arranged in tandem arrays of up to millions of copies. The copies may have very simple sequences, such as the poly(AT) satellite of the crab, *Cancer borealis* (Sueoka and Cheng, 1962), or they may be more complex, such as the 1.4kb and 2.6kb bovine satellite sequences (Streeck, 1982). Satellite sequences are predominantly associated with heterochromatin, though this is certainly not their only location (John and Miklos, 1979). If highly repeated sequences are localised in large tandem arrays, they could only influence gene expression in *trans*. There have been some reports of transcription from satellite sequences (e.g. Varley et al., 1980 and Diaz et al., 1981) though not apparently from heterochromatic locations.

In contrast to highly repeated DNAs, reassociation studies have shown that up to 80% of many eukaryote genomes consists of middle repeated DNA interspersed amongst single copy sequences. This middle repeated sequence component is potentially capable of interacting with genes in *cis* as well as in *trans* and is therefore in a better position to influence the expression of the genome than is the bulk of highly repeated sequences. The interspersal of middle repeated sequences with unique sequences also means that several arrangements are possible, as will be shown in the next section.

1.2 Organisation and Structures of Middle Repeated Sequences

1.2.1 Reassociation experiments

In the genus *Xenopus*, middle repeated DNA has an average size of 300bp and is interspersed amongst unique sequences of about 1000bp in length (Davidson et al., 1973). The middle repeated component in several invertebrate species, on the other hand, can be divided into classes of short (200-400bp) and long (1.5kb) sequences, both of which are interspersed at an average spacing of 2.0kb (Goldberg et al., 1975). Though this '*Xenopus* pattern' of short period interspersion is fairly widespread (Britten and Kohne, 1968), it is not universal. *Drosophila* (Manning et al., 1975) *Apis* (Crain et al., 1976), chicken (Epplen et al., 1978) and water mold (Hudspeth et al., 1977) all have a long period interspersion organisation, involving longer repeats (up to 5kb for *Drosophila*) interspersed with stretches of single copy DNA averaging 5-15kb in length. The genome of the crab, *Geryon quinquedens* contains very little middle repeated DNA and instead it is the highly repeated sequences that are interspersed with unique DNA (Christie and Skinner, 1979). General rules for repeated sequence function cannot therefore be inferred from their genomic organisations alone, since these have been shown to be so diverse.

Function may be inferred for a family of repeated sequences if the copies are structurally homogeneous, the implication being that selection has acted in order to

preserve a sequence-encoded function. That this is not always necessarily so, is clear from the fact that there are mechanisms that can homogenise families of apparently functionless sequences (see Dover, 1982 and section 1.3). Structural homogeneity can be tested in the first instance by measuring the thermal stability of the duplexes formed when a population of denatured, middle repeated sequences is allowed to reassociate. For example, with renatured *Xenopus laevis* middle repeated DNA, it is found that the shorter repeats (200-400bp) melt over a broad range of temperatures averaging 11.5°C below that for native DNA, while the less abundant longer repeats melt at only 1°C below. The longer repeats are therefore more structurally homogeneous than the shorter ones. This distinction between short and long repeats is seen in clam and sea urchin DNAs as well (Galau et al., 1976).

On comparing the genomes of two closely related species, *X. laevis* and *X. borealis*, which share 70% of their unique DNAs, it is found that most repeated sequence families are present in both, but are considerably reduced in frequency in the one species when compared to the other (Galau et al., 1976). Therefore, either copies have been lost or gained since the species separated, or the repeated sequences have diverged sufficiently to appear unrelated under the conditions used. One cannot distinguish between the two possibilities using renaturation experiments alone. However, even though the two *Xenopus* species have almost identical proportions and distributions of repeated sequences as a whole, the fact that

the *amounts* of particular repeated sequences can vary suggests that repetition frequency *per se* cannot be an important functional attribute.

Renaturation experiments can give only an average picture of the structure and organisation of repeated sequences. They cannot show how heterogeneous are the different copies of a particular family. This is an important point to consider when attempting to apply a single functional attribute to the members of a repeated sequence family, for shared function implies that they will have a certain degree of sequence homology. Recombinant DNA technology allows one to analyse individual copies and examine directly the organisation and structure of a defined repeat family.

1.2.2 Studies with cloned repeated sequences

Some repeated sequence families are known to be coding and recombinant DNA clones containing copies can be isolated with respect to their products. For example, clones containing the 18S and 28S ribosomal genes were selected on the basis of hybridisation to ribosomal RNA (e.g. Long and Dawid, 1980). Ribosomal genes represent 20% of middle repeated DNA in *Drosophila*, yet they are arranged in a tandem array (of about 250 copies) like highly repeated sequences. There are approximately 160 copies of the 5S RNA gene in *Drosophila*, also arranged tandemly (Hershey et al., 1977). The 600 transfer RNA genes are distributed in about 30 clusters (Yen and Davidson, 1980) and thus have the

interspersed organisation that is more typical of middle repeated sequences. However, the tRNA gene units (370bp) are similar in size to the middle repeated sequences typical of the '*Xenopus* pattern' rather than to those of the '*Drosophila* pattern'. *Xenopus* tRNA genes are arranged in a number of clusters as well (Clarkson et al., 1973). Other repeated sequences that code for proteins can either be arranged tandemly, as are the histone genes in a number of organisms (Kedes, 1979), or may be dispersed, as are the actin genes (Fyrberg et al., 1980).

Apart from the ribosomal genes, these gene families do not constitute a significant proportion of the repeated sequence complement. They need not be representative of the bulk of middle repetitve DNA and non-coding repeated sequences need to be examined as well.

Clones containing repeated sequences may be selected randomly from a genomic library (e.g. Wensink et al., 1979 and Sun et al., 1984) or strategies may be employed specifically to isolate them (e.g. Gilroy and Thomas, 1983). Repeated sequence clones may come to light in the course of a chromosomal walk (e.g. Spierer et al., 1983) or clones may be selected on the basis of hybridisation to a predominant RNA species, as were the *copia* sequences discovered by Finnegan et al. (1977). Middle repeated sequences show a variety of structures. These structures are summarised in Figure 1.1 and a more detailed description is given below.

Invertebrates

Much of the data for cloned repeated sequences from invertebrates are derived from studies in *Drosophila*. There are at least five main types of middle repeated sequence in this organism, that together make up 12% of the genome.

Copia-*like sequences*. The *copia*-like set of sequences make up half the middle repeated DNA of *Drosophila*, and it is estimated that there are up to 40 families with 30-50 members in each (Young, 1979). The families are not homologous, but share a similar structure - a 3 to 8 kb region is flanked by direct repeats of about 300bp in length. These direct repeats are themselves flanked by smaller (3bp to 17bp) inverted repeats. Finally, the elements are flanked by short direct repeats, thought to be generated by duplication of a host target sequence on insertion. *In situ* experiments show that such *copia*-like sequences are dispersed throughout the euchromatin and are present also in the chromocentre. Some euchromatic locations differ when strains are compared, suggesting that these sequences are mobile within the genome (Young and Schwartz, 1980). One spontaneously derived mutation at the white locus, *white-apricot* (w^a), is due to the insertion of a *copia* sequence (Rubin, 1983). Nomadic elements with structures similar to those of the *copia*-like sequences in *Drosophila* are present in yeast (Eibel et al., 1980), slime molds (Chung et al., 1983) and nematodes (Emmons et al., 1983).

Transcripts of the *copia* sequence are present in the

nuclear and cytoplasmic poly(A)$^+$ RNAs of *Drosophila* tissue culture cells. There are two major transcripts of 5kb and 2kb in size, as well as a more heterogeneous population. The transcripts are also abundant in larval tissues, but absent in adults and embryos (Rubin et al., 1980). Other *copia*-like sequences are also transcribed, if less abundantly (Georgiev et al., 1980).

P-elements. These sequences have short (31bp) inverted terminal repeats. The internal region is variable in length, with a maximum of 2.9kb. Shorter P-elements are derived from deletions of the intact 2.9kb element (Rubin, 1983 and O'Hare and Rubin, 1983). Intact P-elements are thought to encode both transposition and repression of transposition functions. When an egg from a strain that lacks functional P-elements (M strain) receives a sperm from a strain that does possess them (P strain), then all integrated P-elements are potentially capable of transposing through the genome. This can result in a variety of disruptive events, such as the insertion of a P-element into a coding sequence, that are collectively called 'hybrid dysgenesis'. In the presence of an intact P-element, a sequence located between the two inverted terminal repeats is able to integrate into the *Drosophila* genome. This has made possible the construction of vectors that will mediate the transfer of subcloned genes into different *Drosophila* genomes (Rubin and Spradling, 1982).

Foldback (FB) elements. Foldback (FB) elements are also thought to be mobile (Truett et al., 1981). A variable

internal region is flanked by a pair of *inverted* repeats which are not however exactly identical. The structure of the inverted repeats, starting from the distal end, consists of a set of 10bp repeats interspersed amongst unrelated DNA. Progressing inwards, the 10bp repeats expand into 21bp, then 33bp and finally 155bp repeats. At this stage there is a tandem array of these repeats, until the internal region is reached. In some cases, no internal region is present (Potter, 1982). The instability of the mutation *white-crimson* (w^c) is due to the mobility of an inserted FB element that has a 4kb internal region surrounded by the inverted repeats (Levis et al., 1982). Repeated sequences similar to the *Drosophila* FB elements have so far only been described in the sea urchin, *Stronylocentrotus purpuratus* (Liebermann et al., 1983). The 200-400 copies of this 'TU' family of foldback sequences are also dispersed in the genome. Their inverted terminal repeats are each 800bp long and the internal region again is of variable size.

'Clustered and scrambled' sequences. The dispersed, 'clustered and scrambled' repeats are about 1kb long, but are compounds of three or four different smaller units. Different copies contain different arrangements and sometimes different types of the smaller repeats (Wensink et al., 1979). At least one type has different cytological locations in different strains. Such sequences are also distributed throughout 20% of the genome of the slime mold, *Physarum polycephalum* (Peoples and Hardman, 1983).

F elements. The F family of repeated sequences (DiNocera et al., 1983) consists of sequences that are mostly 4.7kb in length. Shorter elements are truncated at their 5' ends, so that all copies have a common 3' end and an accompanying poly(A) tail. There are no internally redundant sequences. The elements are dispersed in the *Drosophila* genome at about 25 euchromatic sites. F elements occupy different sites in different *Drosophila* strains and insertions generate duplications of 8 to 13 bases of the host sequence.

Drosophila melanogaster has at least three times as much repeated DNA as does *Drosophila simulans*, and though most randomly selected repeated sequence clones from a *D. melanogaster* library are also present in *D. simulans*, they are less abundant. Furthermore, whereas in *D. melanogaster* the locations of these repeated sequences are dispersed, in *D. simulans* they tend to be limited to single sites (Dowsett and Young, 1982). *Copia* and 412 sequences are present in *D.melanogaster, D. simulans* and *D. mauritiana* genomes, but not in those of *D. erecta* and *D. yakuba*. Conversely, a number of repeated sequence clones from a *D. erecta* library have been shown to be restricted to *D. erecta* and *D. yakuba*. In all, of 61 repeated sequence clones tested, only 9 were found in all five *Drosophila* species. Two of these were ribosomal DNA clones and the other seven could be either histone or tRNA genes (Dowsett, 1983). A screen of a larger range of *Drosophila* species showed *copia* and 412 homologous sequences

to be present in most, whereas 297 (a *copia*-like sequence) and another repeated sequence (the TIP sequence) were limited to the *melanogaster* subgroup (Martin et al., 1983).

There are no obvious features that are shared by all five classes of repeated elements in *Drosophila*. Most contain mobile sequences, though whether all elements are flanked by small duplications of the host sequence is not known. It is possible that some sequences of the 'clustered and scrambled' type are in fact analogous to elements of the other classes for they have not been examined in the same detail.

Apart from sharing the overall structure, the different families of elements from within a class have little in common. For example, the *Drosophila copia*-like elements do not cross-hybridise and are not equally represented in other *Drosophila* species (Finnegan et al., 1982).

Some of these classes of structures are represented also in vertebrates, which also have repeated elements with structures distinct from those found in invertebrates.

Vertebrates

Copia-*like elements*. The proviral (integrated) form of retroviruses, though not strictly part of the genome, is probably the mammalian equivalent to the *copia*-like sequences of *Drosophila*. They consist of internal coding sequences that are flanked by 500bp direct repeats (or long terminal repeats, LTR). The fact that *copia* elements exist in double stranded

circular form in tissue culture cells (Flavell and Ish-Horowicz, 1981), as do retroviruses, and may be equivalent to virus particles (Shiba and Saigo, 1983), lends further support to the suggestion that *copia* and retroviruses belong to the same general class of repeated sequences. Sequences homologous to the direct repeats of one class of retroviruses, and separate from retroviral internal sequences, are also dispersed in the mouse genome, but is is not known if they are transposable (Wirth et al., 1983). An analogous situation is seen with the insertion sequences (IS) and transposons of bacteria.

P-elements and foldback elements. No repeated sequences analogous to these invertebrate repeated sequence classes have been reported in vertebrate genomes as yet.

'Clustered and scrambled' elements. 'Clustered and scrambled' elements are present in the genomes of the chicken (Musti et al. 1981) and of the rat (Alonso et al., 1983). Three elements from a 1.0kb DNA fragment of the rat genome were sequenced and they show no obvious sequence homology with each other, lending further support to the idea that 'clustered and scrambled' structures consist of adjacent but unrelated repeated sequences.

Much of the middle repeated component of mammalian genomes consists of sequences, like the F-elements of *Drosophila*, that have no gross internal repetition. These have been further subdivided by Singer (1982b) into SINEs

(short interspersed repeated sequences) and LINEs (long interspersed repeated sequences). These are present in the order of 10^5 and 10^4 copies, respectively.

SINEs. The predominant SINE family in primate genomes is the Alu family of repeated sequences (Schmid and Jelinek, 1982). There are at least 300,000 copies of the Alu sequence in the human genome, making up 10% of the total repeated sequence component. Alu sequences are dimers of a 130bp sequence, both of which have poly(A) stretches at their 3' ends. The elements are flanked by the short direct repeats that have been associated with mobility. The sequences are evenly distributed throughout the genome because 90% of randomly selected genomic clones contain Alu members. 10-20% divergence is shown between ten sequenced copies (Deininger et al., 1981). The B1 family of repeated sequences in the mouse is equivalent to the 130bp monomer of the Alu sequence, and members of the B2 family have some homology to the Alu sequence (Georgiev et al., 1982). An Alu probe also hybridised to sequences in the genomes of birds, amphibians and echinoderms (Ullu, 1982).

Other short middle repeated sequences have been identified in human genomic clones (Sun et al., 1984). Two sequenced copies of the so-called 'O' element family had common 5' ends, but one copy extended at its 3' end into a poly(A) tract.

LINEs. The two major types of vertebrate long interspersed repeated sequences are the KpnI and 'MIF-Bam-R'

families of primates and rodents, respectively. KpnI sequences have 60-70% seqeunce homology with part of the rodent LINE sequence (Rogers, 1983 and Singer et al., 1983).

The KpnI sequences show up as a predominant 6kb band when human genomic DNA is digested with KpnI enzyme and separated on an agarose gel and there are about 40,000 copies in this species. No short sequences appear to be repeated within a KpnI element. Some copies may be truncated at their 5' ends but all sequenced copies have a common 3' end as well as a poly(A) tract (DiGiovanni et al., 1983).

The three components of the MIF-Bam-R elements were initially described separately. 'R' repeated sequences are about 500bp in length and there are about 10^5 copies in the mouse genome (Gebhard et al., 1982). Some R copies extend at their 5' end into sequences equivalent to the 'Bam' elements (Fanning, 1982) and some of these extend in turn into sequences equivalent to the Mouse Interspersed Family or MIF-1 elements (Brown and Dover, 1981). The entire MIF-Bam-R element is about 7kb in length and is present in about 10^4 copies. It is suggested that this element can transpose via an RNA intermediate, with transcription starting at the 3' end where the R element is situated. Incomplete transcription would lead to the transposition of sequences lacking the MIF or even the Bam components - in other words, the R sequences only may be transposed (Rogers, 1983). Like the KpnI and Alu elements, all these sequences have a poly(A) tract at their 3' ends.

Non-mammalian LINEs. No repeated sequences equivalent to the mammalian LINEs have been reported in other vertebrate species. However, two other types of middle repeated sequence have been described in *Xenopus*. The 1723 element (Kay and Dawid, 1983) has two pairs of inverted terminal repeats of total length 500bp and an internal region whose length varies from 3kb to 6kb. This variability is due to changes in the number of 185bp tandem repeats that reside in the internal region. Another repeated sequence family consists of two sets of related, and tandemly arranged, 400bp repeated sequences (Carroll et al., 1984). Again, the number of elements in the tandem array can vary. The sequences also share distinct 5' and 3' 1kb flanking regions which contain no internal redundancy themselves.

In summary, some classes of repeated sequences appear not to be shared by both vertebrates and invertebrates, though this may well be because the various genomes have not been exhaustively screened for all their constituent repeated sequences. The classes that are represented by the most abundant sequences will tend to be the first to be detected. Therefore, the apparent differences between vertebrate and invertebrate genomes may simply reflect the relative abundances of the various repeated sequence classes present. Of course, some functional significance may be attached to the fact that one class of repeated sequences is very abundant in on genome and not in another. For example, why are short

interspersed sequences so abundant in mammalian genomes but relatively scarce in *Drosophila*? However, to see how copy number may influence the function of a particular repeated sequence family, one first needs to know if indeed it has a function. The following section discusses the roles that repeated sequences are thought to play in a variety of organisms.

1.3 Possible Functions for Middle Repeated Sequences

The diversity of structures and the dispersed arrangement shown by middle repeated sequence has been related to two contrasting, though not necessarily mutually exclusive, views of their role in the eukaryote genomes. On the first view, one need attribute no function to repeated sequences. Instead, if a sequence is capable of replicating within the genome, and has no effect on the fitness of the phenotype, its numbers will inevitably increase (e.g Orgel and Crick, 1980). Two processes, 'replication and transposition' (Calos and Miller, 1980) and 'unequal crossover' (Smith, 1976) can provide the mechanisms for such behaviour, though the former process would be most applicable to families of dispersed repeats. Any tandemly repeated sequence is potentially capable of undergoing unequal crossover, but possibly only sequences having structures similar to those already discussed are transposable. These two processes, together

with gene conversion (e.g. Klein and Petes, 1981), could be responsible for the homogenisation of a repeated sequence family in a genome. This process of homogenisation and the consequent fixation of a repeated sequence variant in a population have been combined in the term 'molecular drive' (Dover, 1982). Molecular drive aims to explain why there is as little as a tenth less variation between copies of a repeated sequence family within a species than between species. Repeated sequences generated in this way which have no influence on the expression of the genome, and yet employ the genome's replication machinery, are regarded variously as 'selfish' (e.g. Doolittle and Sapienza, 1980), 'parasitic' (e.g. Orgel and Crick, 1980) or 'ignorant' DNA (Dover, 1980).

Alternatively, dispersed repeated sequences would satisfy the conditions required for a model of gene regulation during development (Davidson and Britten, 1973 and 1979). Sets or batteries of genes that are themselves dispersed could be coordinately expressed if they shared common 'sensor' sequences. These 'sensor' sequences would respond to a given cellular signal (such as a hormone-receptor complex) by switching on, or off, a particular battery of genes. In a further speculation, the same authors propose that the apparent mobility of repeated sequences will provide the genome with the plasticity required to alter developmental pathways and so generate new lineages of organisms.

What evidence is there for this model? Much sequence data has accumulated for functionally related genes and some

common sequences have been found to lie 5' to them. However, these shared, repeated sequences are quite short (from 9 to 20 nucleotides - Davidson et al., 1983) and so are not equivalent to the middle repeated DNAs as they are usually defined.

Although only a third of the sea urchin, *Strongylocentrotus purpuratus* unique DNA is interspersed with repeated sequences, this fraction is responsible for almost all of the messenger RNA at the gastrula stage (Davidson et al., 1975). Also, both strands of 80% of the short repeat families, and of 35% of the long repeat families, are represented in total oocyte RNA (Costantini et al., 1978), though not in amounts that are proportional to their frequencies of repetition in the genome. However, these results do not indicate if repeated sequences are directly involved in the production or maturation of transcripts, nor do they show how many members of a particular family are actually transcribed. More specific results are obtained by using cloned DNAs as probes.

Tchurikov et al. (1982) found a 200bp sequence, present in 200 copies in *Drosophila*, that lay 3' to a particular cloned gene, to be present in at least 20 cytoplasmic poly(A)$^+$ RNA species. It was not shown whether the sequence lay at the 3' end of all these RNAs. In the slime mould, *Dictyostelium discoideum* one 300bp repeated sequence hybridises to 1% of poly(A)$^+$ RNA from vegetative cells (Kimmel and Firtel, 1979). The amount of this hybridising RNA species increases five-fold, some five hours after the onset of development into a spore-forming body (Davidson and Posakony, 1982). From the

same organism, Zuker and Lodish (1981) isolated two classes of cDNA clones that contained a copy of another repeated sequence family. Both classes are absent from vegetative cell RNA. One class is transcribed at the onset of development, while the other is not transcribed until 15 hours later. These authors have suggested that the repeated sequences were functionally associated with the developmentally regulated RNAs, though it remains possible that they are transcribed by default if, for example, they were situated within introns.

The most convincing example of a repeated sequence that is functionally connected with a set of genes is the 'homeo box', a 180bp sequence lying at the 3' ends of a number of homeotic genes in *Drosophila* (McGinnis et al., 1984). The repeated sequence is itself part of the translated genes, and homologous sequences have been found in *Xenopus*, mouse and human genomes. It is suggested, however, that the 'homeo box' defines a set of functionally related genes, and is not necessarily required to coordinate expression.

What other functions may be carried out by repeated sequences? The human Alu sequences are transcribed *in vitro* as short transcripts (using RNA polymerase II) or as parts of longer transcripts (using RNA polymerase III) *in vivo* (Haynes and Jelinek, 1981). The Alu consensus sequence shows homology to the rodent 4.5S and human 7S RNA genes (Jelinek and Haynes, 1982). Alu sequences are also **adjacent to** several deletions that can cause thalassaemias (Ottolenghi and

Giglioni, 1982) and hereditary persistence of foetal haemoglobin (Jagadeeswaran et al., 1982), and with other genomic rearrangements (Calabretta et al., 1982).

Repeated transposable sequences are responsible in part for reversible mutations and chromosomal rearrangement in maize. Activator (Ac) sequences are able to transpose themselves to different locations, and can induce the smaller, Dissociator (Ds) sequences to transpose as well.* A Ds element (one of 30-40 copies in the maize genome) was isolated after selecting for reversible mutant Adh genes that had been induced by the Ac-Ds system. This element is 405bp large, with 11bp inverted terminal repeats, and generated direct repeats of 8 nucleotides of the host sequence on insertion (Sutton et al., 1984). Mobile genetic elements are also involved in genome alterations associated with adaptive changes of phenotype in prokaryotes. For example, the inversion of insertion-like sequences is responsible for a) the alternate expression of two types of flagellin genes in *Salmonella* (Sivermann et al., 1979) and b) the alternative expression of genes that determine the different host ranges of Mu phage (Bukhari and Abrusio, 1978).

A number of phenotypic effects can therefore be associated with repeated sequences. However, no clear functions have been assigned to several types of middle repeated sequences, such as the 'scrambled and clustered' repeats of *Drosophila* and chicken, or the large KpnI repeats in humans, even though some are known to be transcribed. In

* B. McClintock (1951). 'Chromosome organisation and genic expression.' Cold Spring Harbor Sym. Quant. Biol. 16, 13-47.

investigating the possible function of a repeated family, one must consider most of the copies, for one copy only may actually be required. Pseudogenes, for example, are copies of functional genes that cannot themselves produce functional mRNAs, and so are probably redundant. Ideally then, each copy of a repeated sequence family should be isolated and analysed. Obviously, it is desirable to deal with a repeated sequence family for which there are few copies, making the Alu family for example, with its 300,000 copies per haploid human genome, a poor candidate. Though it can be argued that low repetition frequency families may have functions quite distinct from higher repetition frequency families, the aim is to provide a complete description for any repeated sequence family. Secondly, repetition frequency alone is not a suitable criterion for classifying a repeated sequence family, for, as mentioned earlier, families that are abundant in one species may be in low copy number in a close relative.

This thesis therefore sets out to describe one particular repeated sequence family whose properties suggest that it may be an ideal system for analysing possible repeated sequence function. As will be described in more detail in the next section (section 1.4), this family is present in many species (Singh et al., 1980b) and is in low copy number in *Drosophila*. Various authors have suggested that this repeated sequence family is intimately involved in sex determination and may even be the sex determiner itself (e.g. Epplen et al., 1983b

and Jones, 1983). Whether this is true or not, the elucidation of the role of these repeated sequences is most likely to come from a study in an organism, such as *Drosophila melanogaster*, where the advantage of having to deal with only a few dispersed copies can be combined with the extensive available genetic knowledge.

1.4 The Garden of Eden (GOE) Family of Conserved and Dispersed Repeated Sequences

1.4.1 A class of snake satellite sequences that is associated with sex chromosomes.

In some snake species, as in birds, the female is the heterogametic sex and is designated ZW, while the male is designated ZZ. A satellite DNA fraction, present in the female of the Colubrid snake, *Elaphe radiata* was shown by *in situ* hybridisation to be localised almost exclusively on the W chromosome (Singh et al., 1976). This satellite DNA also hybridised *in situ* to the W chromosomes of other snake species. Species of the more primitive order, Boidae do not have morphologically distinct sex chromosomes and here the *E. radiata* satellite DNA hybridises equally to all chromosomes (Singh et al., 1980a). The ZW sex chromosome systems of birds and snakes and the XY sex chromosome system of mammals are thought to have evolved by the progressive rearrangement of one of the progenitor sex chromosome homologues. As this has the effect of preventing crossover during meiosis, the two chromosomes become genetically isolated and a chromosome based mechanism for sex determination can develop (Ohno, 1967). An alternative mechanism was proposed by Singh et al. (1976). They suggested that this snake satellite DNA was involved in the heterochromatinisation of one of the sex homologues, leading eventually to the genetic isolation of the two chromosomes. The same workers then isolated a minor satellite

DNA fraction from the female of the Elapid snake, *Bungarus fasciatus* (Singh et al., 1980b). This satellite DNA, called Bkm, was shown by analytical centrifugation to be virtually absent in males. Also, like the *E. radiata* satellite, Bkm hybridised predominantly to snake W chromosomes *in situ*. Bkm DNA hybridises also to the genomes of birds, mammals and *Drosophila* as well as to simpler eukaryotes, such as slime molds (Jones, 1983). Sequences homologous to Bkm are also predominantly located on the W chromosome of the Japanese quail. The genomes of male and female mice show no quantitative differences in hybridisation with Bkm DNA. However, Bkm DNA does hybridise specifically to large (>2.0kb) Alu I restriction enzyme fragments of the male but not to the corresponding fragments of the female. This 'male pattern' of hybridisation was also found in sex reversed (XX_{sxr}) mice, which are genotypically female but phenotypically male (Jones and Singh, 1981). Using a Bkm probe as a Y chromosome 'tag', *in situ* hybridisations showed this XX_{sxr} condition to be due to the translocation of a Bkm-containing portion of the Y chromosome to an arm of one of the X chromosomes (Singh and Jones, 1982). This sex-associated, differential hybridisation of Bkm suggested to the authors that Bkm DNA contains sequences that are intimately involved in sex determination.

1.4.2 Demonstration that only one sequence component of the snake satellite DNAs is preserved in other eukaryote genomes.

Epplen et al., (1981) isolated a DNA clone from a library constructed from the female-specific satellite of *E. radiata*. This clone also hybridised specifically to the large Alu I restriction enzyme fragments of the male mouse. Some differences in the hybridisation pattern were seen between male and female human genomic DNAs, though these could have been due to restriction enzyme site polymorphism. Sequencing of part of this snake clone (designated pErs5) revealed a 150bp stretch of contiguous GATA and GACA tetranucleotides (Epplen et al., 1982).

In situ hybridisation of Bkm DNA to *Drosophila* polytene chromosomes showed an intense signal at region 19F-20AB, which lies on the X chromosome, near the euchromatin-heterochromatin junction (Singh et al., 1980b). A number of recombinant clones were then isolated from a *Drosophila melanogaster* genomic library on the basis of hybridisation to Bkm DNA. These clones hybridised not only to the 19F-20AB region, but also to additional euchromatic sites. One possible explanation for this is that Bkm-like sequences are situated in dispersed locations, but only at the 19F-20AB region are they in sufficient quantities to be detected by the Bkm probe *in situ*. The cloned DNAs, though containing Bkm-related sequences, would hybridise to dispersed sequences that were homologous to other sequences contained in the *Drosophila* clones.

Singh et al. (1984) sequenced the Bkm-like regions from two of the *Drosophila* clones and from one mouse genomic clone. As in the case of the snake Bkm-related clone, these three clones also contained regions rich in poly(GATA) sequences, though not many GACA tetranucleotides were apparent. Poly(GATA) is the most obvious element shared by these various Bkm-related clones, though it need not be the only sequence component present in the snake satellite DNAs.

1.4.3 Transcription of the poly(GATA) sequences.

Poly(GATA) sequences are transcribed *in vivo*. The poly(GATA) region of the snake satellite clone hybridised to several fragments in both male and female mouse liver RNA. A mouse cDNA clone was isolated on the basis of hybridisation to the snake (pErs5) clone and this also was found to contain poly(GATA) sequences (Epplen et al., 1983a). Singh et al. (1984) used one of the *Drosophila* clones to probe the RNA from various tissues from male and female mice. Brain tissue RNA from both sexes gave the same pattern of hybridisation. However, liver RNA showed a male-specific pattern of hybridisation. This contrasts with the results of Epplen et al. (above), though this could be explained by the fact that the latter authors used total cellular liver RNA, while Singh et al. used poly(A)$^+$ liver RNA. Poly(GATA) sequences were also reported to be transcribed from *Xenopus laevis* lampbrush chromosomes (Epplen et al., 1983b).

A sequence similar to those described above was recently

isolated from a rat genomic library (Alonso et al., 1983). A 200bp region containing 32 GATAs lies in a 1.3kb fragment along with two other different repeated sequences. Each of these repeated sequences was shown to be transcribed *in vivo*. It is not known if this cluster of repeats is derived from the Y chromosome.

Although the snake satellite DNAs probably contain several sequence components, that component that is preserved in other eukaryote genomes is the poly(GATA) sequence. However, the terms that have been used in the literature to refer to this component, such as Bkm and 'sqr' (for *simple quadruplet repeats*), are not very specific. As this family of repeated poly(GATA) sequences was initially isolated from a snake species, is present in several diverse eukaryote genomes *and* is associated with sex chromosomes, it is referred to here as the Garden of Eden (GOE) family of repeated sequences.

1.5 Strategy for investigating the function of GOE sequences

What strategy should be used to determine whether the GOE family of repeated sequences does indeed have a role in the expression of the genome, especially sex determination? Except where the sequence codes for a structural molecule, such as ribosomal RNA or a protein, one is initially limited to analysing the nucleotide sequence. Even if a family of

repeated sequences was found to be involved in the coordinated expression of a set of genes, the mechanism of action would have to be encoded in the nucleotide sequence. In the case of GOE, the clues that it may have a function are a) its presence in a diverse range of organisms and b) its apparent association with sex determination.

For several nucleotide sequences to have the same function implies that they are to some degree homologous. Homology is usually measured as the percentage of positions that have identical nucleotides in two or more sequences. The question then is, what minimum sequence similarity would indicate that the members of a family of repeated sequences share a function? This will depend on the actual function encoded. For example, a family of protein-coding genes must at least conserve most of the first two nucleotides in each codon. A family of sequences that induce a particular configuration in duplex DNA, on the other hand, may tolerate less homology and still maintain the same structure. However, a high degree of homology need not imply involvement in the expression of the genome, for apparently functionless or 'ignorant' DNAs, such as the ribosomal spacer sequences in *Drosophila*, can maintain sequence homology via the processes that underly molecular drive (Dover, 1982).

The apparent conservation of GOE sequences that is seen between species has been defined by hybridisation experiments, usually with Bkm DNA as a probe. However, if GOE is composed predominately of GATA sequences, one should expect hybrid-

isation of a GOE copy to any sequence rich in GATA, even though other parts may be unrelated. Therefore, the question of conservation of the GOE sequence needs to be analysed at the nucleotide level.

GOE sequences may have the same function in genomes from different species, but some variation might be expected because different species have different genetic backgrounds. Therefore, one needs to compare the copies from within one genome, so that, on the assumption that GOEs are functional, an estimate can be obtained of the amount of variation that is tolerable for this particular family. If the variation can be shown to be essentially random (that is, it is not localised or it is not limited to certain nucleotides) then this would be good evidence that GOE sequences have no function. Though one could argue that only one GOE copy may be functional, so that the remainder can mutate at random, one is then dealing, functionally-speaking, not with a repeated sequence, but with a unique sequence.

The aim of this thesis is to isolate and sequence a majority of the GOE sequences from the *Drosophila melanogaster* genome and compare them at the nucleotide level. From this, an estimate of the intragenomic variation of GOE sequences can be obtained. The possible functions that have been proposed for the GOE family of middle repeated sequences can then be examined in the light of these data.

Figure 1.1

Diagramatic representations of selected classes of repeated sequence families that are discussed in the text. Distinct and/or repeated structures within each element are fully or partly shaded. Diagrams are not to scale.

Drosophila

1) *copia*-element (Finnegan et al., 1977).
2) P-element (O'Hare and Rubin, 1983).
3) Foldback (FB) element (Potter, 1982).
5) 'Clustered and scrambled' elements (Wensink et al., 1979).
4) F element (DiNocera et al., 1983).

Yeast

1) Ty 1 transposable element (Eibel et al., 1980).

Xenopus

1) TU1 element (Liebermann et al., 1983).
2) PR elements (Carroll et al., 1984).
3) 72bp tandem repeats (Spohr et al., 1981).

Human

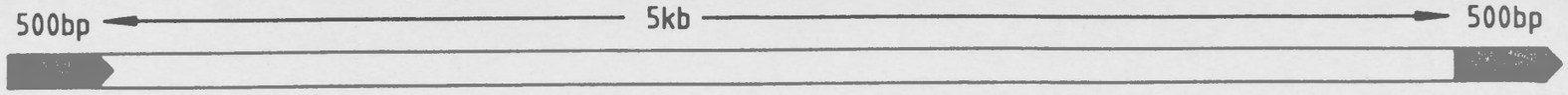
1) Alu family (e.g. Jelinek and Haynes, 1983).
2) 'O' family (Sun et al., 1984).
3) 'K' family (Sun et al., 1984).
4) Kpn I family (DiGiovanni et al., 1983).

Mouse

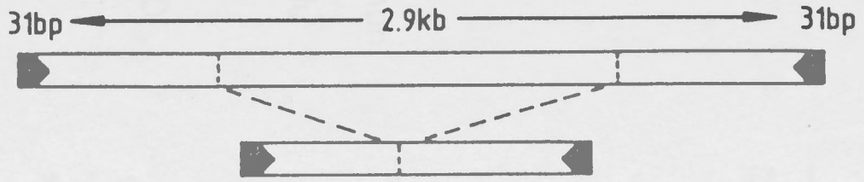
1) MIF, Bam and R families (Brown and Dover, 1981, Fanning, 1982 and Gebhard et al., 1982).
2) PR family (Kominami et al., 1983).
3) Retroviruses (e.g. Varmus, 1983).
4) LTR-IS (Wirth et al., 1983).

Drosophila

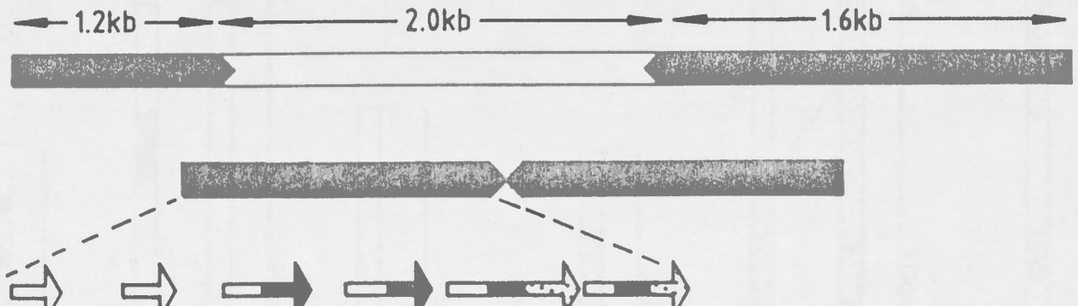
1. Copia



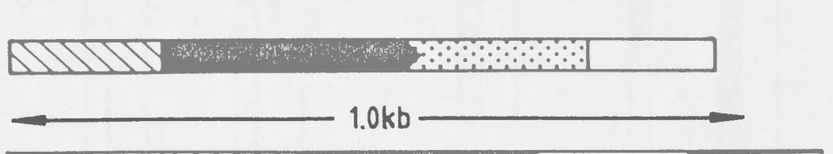
2. P-element



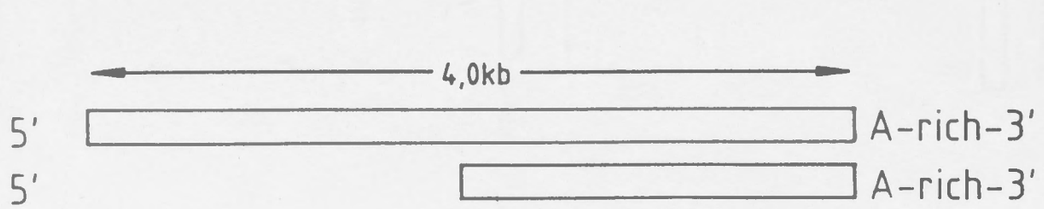
3. Fold back (FB)



4. Clustered and Scrambled

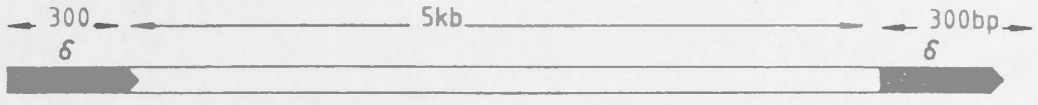


5. F



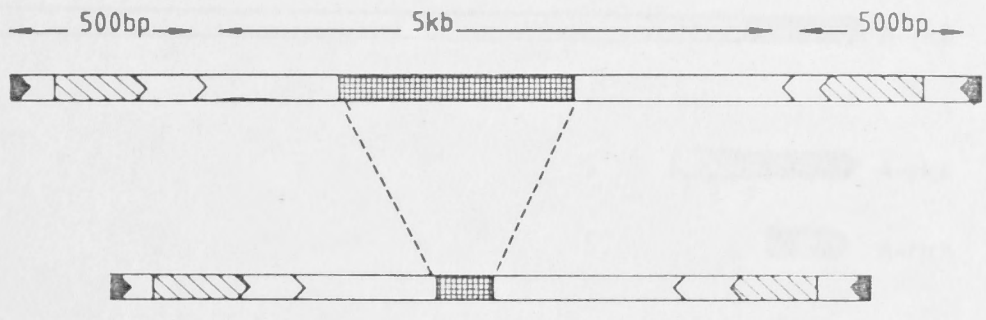
Yeast

1. Tyl

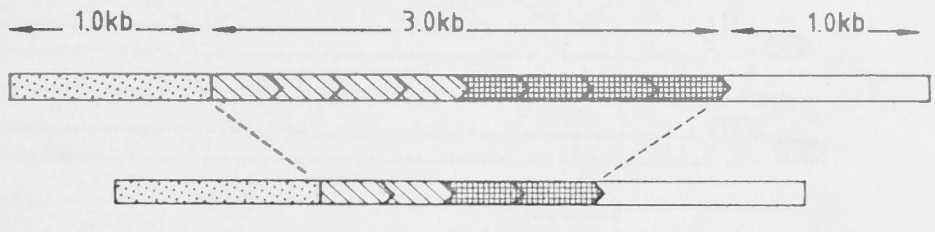


Xenopus

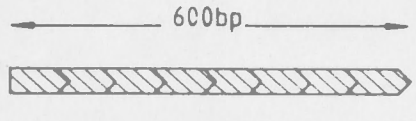
1.



2.



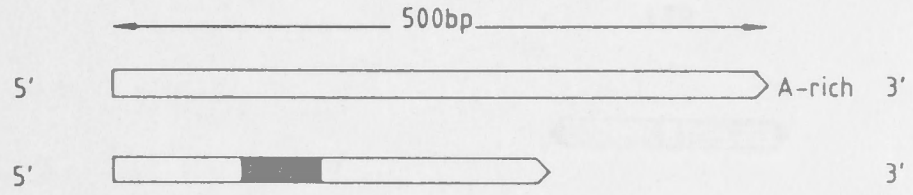
3.



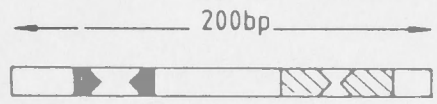
Human

1. Alu

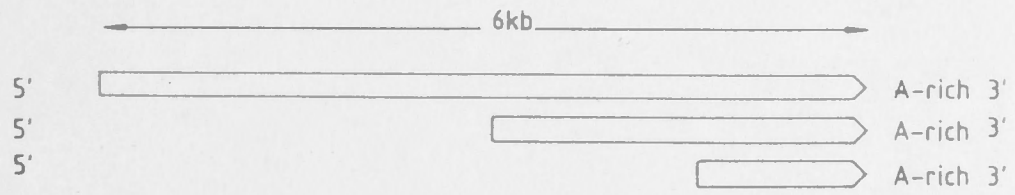
2. 'O'



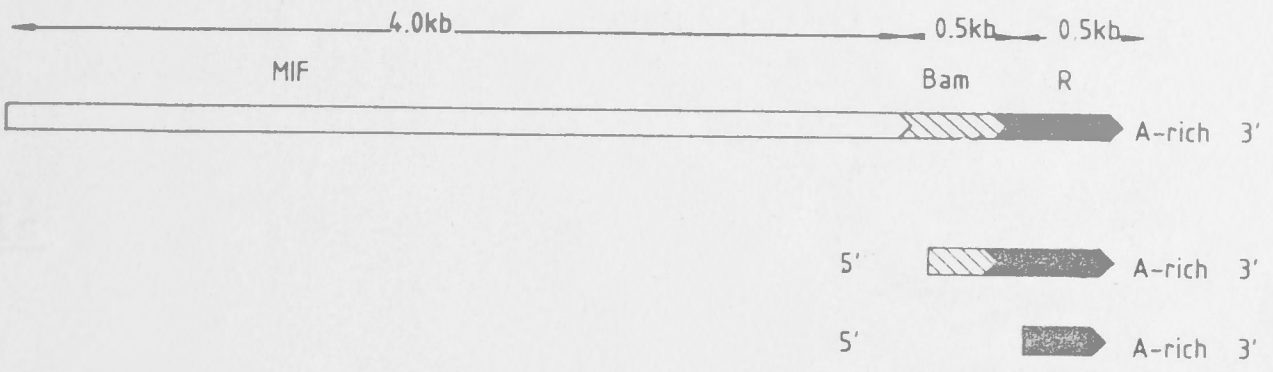
3. 'K'

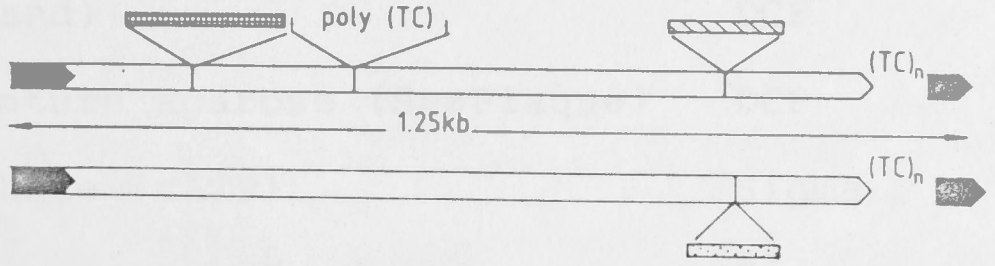


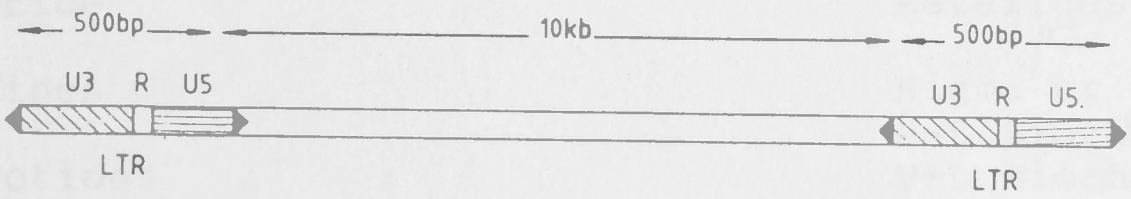
4. KpnI



Mouse







4. LTR-IS

Chapter 2

MATERIALS AND METHODS

2.1 Materials

2.1.1 Chemicals and Reagents

Agarose (SeaKem brand)	DCF
Low melting temperature agarose (SeaPlaque)	DCF
Adenosine triphosphate (rATP)	Sigma
Ampicillin	Sigma
Acrylamide	BioRad
Bactotryptone	Difco
Bovine Serum Albumin	Sigma
Caesium Chloride	Metallgesellschaft
Deoxynucleotides	Sigma or P-L Biochemicals
Dideoxynucleotides	P-L Biochemicals
alpha-^{32}P-dATP and alpha-^{32}P-dCTP	Amersham, NEN or Bresa
Ficoll	Pharmacia
Isopropyl-beta-thiogalactopyranoside (IPTG)	Sigma
N,N'-methylene bisacrylamide (Bis)	BioRad
Nitrocellulose (0.45 m)	Schleicher & Schuell
Phenol	Wako Chemical Industries
Polyvinyl pyrrolidone (PVP)	Sigma
Sephadex G-50	Pharmacia

Spectinomycin	Upjohn
TEMED (N,N,N',N',-tetramethyl ethylenediamine)	BioRad
Tetracycline	Sigma
5-Bromo-4-Chloro-3-Indolyl-beta-D-galactopyranoside (X-gal)	Sigma
X-ray film (XS-5 & XRP-1)	Kodak
Yeast extract	Difco

All other chemicals and reagents used were analytical grade.

2.1.2 Enzymes

The following restriction endonucleases were prepared by J. Blok and A. McKenzie:

Bam HI, Eco RI, Hind III and Pst I.

All other restriction enzymes were obtained from New England Biolabs or Amersham.

E. Coli DNA polymerase I (Klenow fragment) was obtained from New England Biolabs or New England Nuclear.

BAL 31 exonuclease was obtained from New England Biolabs.

T4 DNA ligase was a gift from A. McKenzie.

Calf intestinal alkaline phosphatase (CIAP) was obtained from Collaborative Research.

2.1.3 Bacterial strains and DNA vectors

RR1: F$^-$, hsd S20 (r_B^-,m_B^-), ara-14 proA2,lacY1, galK2, rpsL20 (Sm^r), xyl-5, mtl-1, supE44, lambda$^-$ (Bolivar et al., 1977).

LE392: F$^-$, hsd R514 (r_K^-, m_K^-), supE44, supF58,lacY1, galK2, galT22, metB1, trpR55, lambda$^-$. (Murray et al., 1977).

JM103: Delta(lac pro), thi, strA, supE, endA,sbcB, hsdR$^-$, F'traD36, proAB, lacI^q, Z DeltaM15 (Messing et al., 1981).

RR1 and JM103 cells were grown in LM broth (10g bactotryptone, 1g yeast extract, 10g NaCl, 2g maltose and 0.2g $MgCl_2.2H_2O$ per litre distilled water).
LE392 cells were grown in NZCY broth (10g NZamine, 5g yeast extract, 5g NaCl, 2g $MgCl_2.2H_2O$ and 1g casamino acids per litre of distilled water).

DNA fragments were cloned into the plasmid vectors pBR322 (Bolivar et al., 1977) and pBR328 (Soberon et al., 1980) and into the bacteriophage vectors M13mp8 and M13mp9 (Messing and Vieira, 1982).

2.2 DNA preparations

2.2.1 Genomic DNA

Genomic DNA from either embryos, adult heads or ovaries, was prepared by G. Miklos and others, as described in Miklos et al., (1984).

Briefly, heads or embryos were homogenised in TE buffer (10mM Tris-HCl, pH8.0, 1mM EDTA) and lysed with Sarkosyl (final concentration of 3%). Caesium chloride was added to the lysate to a final concentration of 1g/ml followed by ethidium bromide (0.6mg/ml). DNA was equilibrated by centrifugation at 45,000 rpm for 40-44 hrs. The DNA band was visualised by UV light (wavelength 254nm) and withdrawn, following side puncture, with a 19 guage needle. The ethidium was removed by repeated mixing with isopropanol, and the aqueous phase dialysed against 2 changes of 2 litres of TE at 4°C. The dialysate was extracted with phenol:chloroform (1:1) and precipitated in 0.3M sodium acetate (NaAc) and 2.5 volumes of ethanol. Pelleted DNA was resuspended in 100-200ul 1xTE.

2.2.2 Lambda phage DNA

$5x10^3$ pfu of bacteriophage in SM (0.05M Tris-HCl, pH7.4, 0.1M NaCl, 0.01M $MgSO_4$ and 0.01% gelatin) were adsorbed to 200ul of stationary phase LE392 cells in the presence of 200ul Mg.Ca (0.01M $MgCl_2$, 0.01M $CaCl_2$) for 10 minutes at 37°C. This was added directly to 200ml of NZCY broth and incubated overnight at 37°C. Phage particles were pelleted by spinning

the supernatant at 11,000 rpm in a GSA rotor for 3 hours. The pellet was resuspended in 4ml of ice-cold TE plus 3.4g CsCl and the suspension spun in a SW65 rotor at 35,000 rpm for 16 hours. The equilibrated phage band was withdrawn by side puncture through a 19 guage needle.

DNA was extracted as follows: to 100ul aliquants of the phage suspension were added 10ul of 2M Tris-HCl, pH8.5 and 100ul of deionised formamide. After standing at room temperature for 2 hours, 100ul of distilled water and 600ul of ethanol were added and the tube gently inverted several times and the DNA pelleted by a 5 minute spin in an Eppendorf bench microfuge. Each pellet was resuspended in 50ul TE and these samples used directly for analytical and preparative purposes.

2.2.3 Supercoiled plasmid DNA

Cultures of bacterial strains containing the desired recombinant plasmids were grown to stationary phase in 5ml of LM broth containing antibiotic (ampicillin at 50ug/ml or tetracycline at 20ug/ml, depending on which resistance marker was still intact. Recombinants in the Eco RI site of pBR322 or pBR328 were consistently grown in medium containing ampicillin). 1ml of the stationary phase culture was added directly to 200ml of LM broth plus antibiotic and incubated at 37°C for 4-5 hours until the optical density at 640nm ($O.D._{640}$) had reached 0.4-0.6. Spectinomycin to 250ug/ml was added as solid to the culture and incubation allowed to continue for a further 16-18 hours.

Bacterial cells were harvested and resuspended in 1.6ml of 25% sucrose solution in TE, to which was added 0.4ml of lysozyme (5-10mg/ml). After mixing and sitting on ice for 5 minutes, the lytic reaction was halted with 0.7ml of 0.25M EDTA. Cells were lysed by the addition of 2.4ml of lytic mix (0.5% Triton-X, 0.05M Tris-HCl, pH8.0 and 0.06M EDTA) and vigorous shaking. Cell debris was pelleted by spinning at 16,500 rpm for 40 minutes. To 5ml of the supernatant were added 5g CsCl and 0.3ml ethidium bromide (10mg/ml). The same conditions of ultracentrifugation as for the genomic DNA preparation were used and the lower, supercoiled DNA band was withdrawn. Ethidium extraction and dialysis were as before, but the dialysate was not extracted with phenol:chloroform. Usually, 1-2ml of DNA solution, containing 200-500ug of DNA, was obtained in this manner. These samples were used directly for analytical and preparative enzyme digestions.

2.2.4 M13 double stranded DNA (replicative form)

10^5 pfu of single stranded infective form of M13, stored in SM, were adsorbed to 200ul of stationary phase JM103 cells in the presence of 200ul Mg.Ca, for 10 minutes at 37°C. This was added directly to 1 litre (or 200ml) of LM broth and incubated for 16-18 hours at 37°C. Double stranded DNA was extracted from the harvested cells as for supercoiled plasmid DNA.

2.2.5 M13 single stranded DNA (infective form)

Single white or clear plaques were picked with applicator sticks and inoculated into 2ml of LM broth in Falcon tubes. After 5-6 hours incubation, 1.5ml of the culture was transferred to an Eppendorf tube and the cells pelleted. 1.0ml of the supernatant was withdrawn and added directly to 220ul of 25% polyethylene glycol 6000 (PEG), 2.5M NaCl, vortexed and left at 4°C overnight. The PEG.DNA precipitate was spun for 5 minutes in an Eppendorf microfuge, the supernatant removed and the pellet resuspended in 100ul of TE. After two rounds of phenol:chloroform extraction, DNA was precipitated in 0.3M NaAc and 500ul ethanol at -70°C (alternatively, phage stocks were prepared by adding 20ul of SM to 20ul of the phage suspension). DNA pellets were consistently resuspended in 20ul of distilled water and these samples used directly for sequencing reactions (see section 2.9).

2.3 Enzyme reactions

2.3.1 Restriction endonuclease digestion of nucleic acids

Most single restriction endonuclease digestions and all digestions employing more than one enzyme were carried out in TA buffer (33mM Tris-HCl, pH7.9, 66mM potassium acetate, 10mM magnesium acetate and 0.5mM DTT). For analytical purposes, two units of the required enzyme(s) were added to 0.5-1.0ug of

DNA in 22ul of TA buffer, and digestion allowed to proceed at 37°C (67°C for Taq I) for two hours. Digestions with Eco RI enzyme only were carried out in Eco RI buffer (100mM Tris-HCl, pH7.5, 50mM NaCl and 5mM $MgCl_2$) and those with Hae III enzyme only were carried out in Hae III buffer (6mM Tris-HCl, pH7.4, 50mM NaCl, 6mM $MgCl_2$ and 6mM beta-mercaptoethanol). Reactions were stopped by the addition of 5ul of sample dye, (30% sucrose, 0.09% bromophenol blue, 50mM EDTA). Preparative digestions (>10ug DNA) were carried out in final volumes of 50ul or 100ul for two hours, scaling up the amount of enzyme used. Reactions were stopped by the addition of 1/20th volume 0.25M EDTA or by immediate extraction with phenol: chloroform. DNA was precipitated in 0.3M NaAc and 2.5 volumes of ethanol at -70°C.

2.3.2 BAL 31 exonuclease digestion*

25ug of plasmid DNA were linearised by digestion with restriction endonuclease for 2hrs, as in section 2.3.1. The reaction solution was made to 100ul in BAL 31 buffer (20mM Tris-HCl, pH8.0, 0.6M NaCl, 12mM $CaCl_2$, 12mM $MgCl_2$ and 1mM EDTA) without removal of the restriction endonuclease reaction components and equilibrated to 37°C before addition of 2 units of BAL 31 exonuclease. 25ul aliquots were removed at four time points and the reactions stopped by immediate extraction with phenol:chloroform. Precipitated and resuspended DNAs were digested with Eco RI enzyme, phenol:chloroform extracted and ethanol precipitated again prior to ligation to the Hinc II and Eco R1 sites of the M13

* Legerski, R.J., J.L. Hodnett, H.B. Gray (1978). Nucleic Acids Research 5, 1445.

vector. Routine agarose gel electrophoresis was carried out to monitor the extent of both BAL 31 and Eco RI digestions.

2.3.3 Dephosphorylation of vector DNA

After vector DNA had been digested completely with restriction endonuclease in TA buffer, 1/10th volume of 1M Tris-HCl, pH10.0 and 1.0 unit of calf intestinal alkaline phosphatase were added. The dephosphorylation was allowed to continue at 37°C for one hour, before two extractions with phenol:chloroform were carried out. After precipitation in 0.3M NaOAc and 2.5 volumes ethanol, the vector DNA was resuspended in distilled H_2O to a final concentration of 0.5ug/ul.

2.3.4 Ligations

20ul of digested plasmid or phage DNA (0.25-0.50ug/ul) together with 5ul of vector DNA were made up to 30ul in Hae III buffer plus 3ul of 10mM ATP (or 1ul 10mM ATP for blunt-ended ligations). Ligation was carried out in the presence of 1.5 units of T4 DNA ligase overnight at 4°C.

2.3.5 High Specific Activity Radioactive DNA Probes

a) *Random priming (Whitfeld et al., 1982).* 5ug of DNA in 20ul of Hae III buffer were digested in the presence of 1 unit of Hae III enzyme for ½ hour at 37°C. 2ul of 'random primers' (prepared by treating herring sperm DNA with DNase 1 and fractionating through a DEAE-Sephadex G-50 column, to obtain fragments of 8-20 nucleotides in length) were added and the

DNAs denatured by boiling for 2 minutes and cooling rapidly in ice. The polymerisation reaction took place in a final volume of 30ul in the presence of 1mM dGTP, dATP and dTTP, 3-5ul of alpha-^{32}P-dCTP (alternatively, if alpha-^{32}P-dATP was used, then dATP was replaced with 1mM dCTP) and 1 unit of DNA polymerase I (Klenow fragment) at 37°C for 45 minutes. To remove unincorporated nucleotides, the reaction mix was made up to 100ul in distilled H_2O and laid on top of a spin column of 1:1 Sephadex G-50 in TE, and spun for 5 minutes at 2000 rpm. The collected volume of labelled DNA was 1-2ml. The activity of 20ul was measured using a Geiger-Mueller tube. Specific activities of 10^7-10^8 cpm/ug were obtained.

b) *Specific priming from M13 templates.* 15ul of single stranded M13 recombinant DNA (prepared as described in 2.2.5) in RT buffer (50mM Tris-HCl, pH8.3, 20mM KCl, 7mM $MgCl_2$, 1mM EDTA and 10mM DTT) plus 2.0ul of 1uM 17-mer sequencing primer were heated at 65°C for 2 minutes and allowed to equilibrate at 37°C for 20 minutes. Polymerisation was carried out in the presence of 1mM dGTP, dCTP, dTTP and 4ul alpha-^{32}a-dATP and 1 unit of DNA polymerase I (Klenow) for 30 minutes at 37°C. The reaction mix was then made up to 100ul and immediately frozen, prior to use as a probe. Unincorporated nucleotides were not removed.

c) *End-labelling*.* The products of standard analytical restriction enzyme digestions (section 2.3.1) were extracted with phenol:chloroform and precipitated in 0.3M NaOAc and ethanol. Pelleted DNA was resuspended in 18ul of Hae III

* This procedure is necessarily restricted to the digestion products of those restriction enzymes that generate a 5'-extension.

buffer. Radioactive and non-radioactive nucleotides and DNA polymerase I were added as in section 2.3.5a. The polymerisation reaction was stopped by addition of 5ul of sample dye (section 2.3.1). Aliquots were separated on analytical agarose or polyacrylamide gels.

2.4 Bacterial transformations

Competent RR1 and JM103 cells were prepared by harvesting log phase cells ($O.D._{640}$= 0.4-0.6) and resuspending in 10ml of ice-cold 0.1M $MgCl_2$ per 100ml of broth used. Cells were spun down and resuspended in 20ml of 0.1M $CaCl_2$. After sitting on ice for one hour, the cells were again spun down and resuspended in 1ml of 0.1M $CaCl_2$. 1ml glycerol was added to competent RR1 cells and these were stored at -70°C in aliquants of 500ul. Competent JM103 cells were prepared fresh each time from 100ml cultures.

Overnight ligation reactions (section 2.3.4) were made up to 100ul with distilled H_2O and 25ul aliquots added to 200ul of competent RR1 cells in the case of plasmid recombinants, and 200ul of competent JM103 cells in the case of M13 recombinants. After sitting on ice for 45 minutes, cells were heat shocked by placing at 45°C for 2 minutes.

Transformation of JM103 cells was also carried out with 1ul of single stranded M13 DNA (prepared as in 2.2.5)

replacing the ligation mix.

a). To transformed RR1 cells 1ml of LM broth was added, and the cells then incubated at 37°C for ½ hr. 0.3ml aliquants were then poured onto LM plates containing 1.5% agar and antibiotic at concentrations equivalent to those used for culturing bacterial strains harbouring plasmids. These were incubated at 37°C overnight.

b). 200ul of transformed JM103 cells was added to a Falcon tube containing 200ul of stationary phase JM103 cells, 7.5ul 0.2M IPTG and 7.5ul 10% X-gal (10ug in 100ul of dimethylformamide). This was mixed with 4ml of molten top agar (0.7% agarose in LM) and poured onto LM plates (1.5% agar only). Incubation was at 37°C overnight.

2.5 Selection of recombinants

2.5.1 Plasmids

Individual colonies were transferred with an applicator stick to duplicate LM plates (one containing tetracycline at 20ug/ml and the other ampicillin at 50ug/ml) and incubated overnight at 37°C. Colonies failing to grow on one plate were picked from the other, suspended in 23ul of colony lysis mix (20ul TE, pH8.0, 1ul lysosyme (10mg/ml) 1ul RNase A (1mg/ml) and 10mM EDTA) and sat on ice for ½ hr.
(Recombinants into the Eco R1 site of pBR322 could not be selected on the basis of antibiotic sensitivity and so

randomly chosen colonies were treated directly with the colony lysis procedure). 5ul of colony dye (25% glycerol, 5% SDS, 0.1% bromophenol blue, 80mM Tris-HCl, pH7.8, 10mM NaAc and 2mM EDTA) were added and the suspensions heated at 65°C for 5 minutes. Suspensions were vortexed vigorously for 1 minute before loading samples onto a horizontal agarose gel. Super-coiled pBR322 DNA was used as a standard.

2.5.2 M13 recombinants

M13 recombinants were selected on the basis of producing white or clear plaques, because the galactosidase gene has been interrupted by insertion of foreign DNA into the cloning region. Up to 20 plaques were picked with applicator sticks, single stranded DNA prepared as described in section 2.2.5, and lul samples spotted onto gridded nitrocellulose filters. The filters were baked at 80°C in a vacuum oven and probed with the appropriate DNA to detect the desired inserts.

2.5.3 Phage plaque screening

For the screening of lambda libraries and of some M13 plaques, the plaque hybridisation method of Benton and Davis (1977) was employed. Plate replicas were made onto pre-cut, circular 0.45um nitrocellulose filters. Phage were lysed by placing the filters in Blot 1 solution (0.5M NaCl, 0.5M NaOH) for 10 minutes, followed by 10 minutes in Blot 2 solution (1.5M NaCl, 0.5M Tris-HCl, pH7.4) and a final soaking in 2xSSC, prior to baking at 80°C in a vacuum oven.

After probing these filters, positive signals on the autoradiograph were lined up with the corresponding plaques on the source plate. A further two rounds of plaque purification were required to isolate single recombinants from the genomic libraries.

2.6 Gel Electrophoresis of Nucleic Acids

2.6.1 Agarose gels.

Analysis of restricted DNA fragments was carried out on horizontal 1.0% agarose slab gels of dimensions: 16cm by 17cm by 1cm. Separation was achieved by electrophoresing samples in Tris-acetate buffer (0.08M Tris-HCl, pH7.8, 0.01M NaAcetate and 2mM EDTA) containing ethidium bromide at 1ug/ml, for 16-18 hours at 1.5-2.5V/cm, depending on the degree of separation required. Migration was monitored by the position of the BPB dye present in each sample. Lambda c1857 DNA, digested with Hind III, provided size markers. DNA bands were visualised by transmitted short wave UV light (254nm).

Preparative DNA samples were separated in 1.0% low melting point (SeaPlaque) agarose horizontal gels (9cm by 5.5 cm by 0.5cm) in Tris-borate buffer (0.1M Tris-HCl, pH8.3, 0.08M boric acid and 2.5mM EDTA) for 2 hours at 3-4V/cm. After staining the gel with ethidium bromide (1ug/ml), DNA was visualised by reflected long wavelength UV light (336nm).

2.6.2 Polyacrylamide vertical gels

Small DNA fragments (50 to 500 bp in size) were separated on vertical 10% polyacrylamide gels (10g acrylamide, 0.5g bis, 500ul 10% $(NH_4)_2 \cdot S_2O_8$ and 30ul TEMED) of dimensions 14cm by 16cm by 1.5mm, in Tris-borate buffer for 3-4 hours at 12-15V/cm.* Gels were stained in ethidium bromide (1ug/ml) and DNA visualised with transmitted UV light (254nm).

Gel patterns were recorded on Polaroid Type 55 Land Film.

2.6.3 Polyacrylamide sequencing gels

Sequencing reactions were separated on thin denaturing (7.0M urea) polyacrylamide gels. For 10% gels, 5g acrylamide, 0.25g bis-acrylamide, 50mg ammonium persulphate and 25g urea were dissolved in a final volume of 50ml of Tris-borate buffer and filtered through 3MM paper. Immediately after addition of 15ul of TEMED, the solution was poured into a gel mould of dimensions 34cm by 36cm by 0.3mm. 8% gels, made using 4g acrylamide and 0.2g bis, were of dimensions 34cm by 100cm by 0.3mm. Also, some reactions were separated on 8% 'wedge' gels of dimensions 36cm by 20cm, whose thickness decreased from 0.6mm at the bottom to 0.2mm at the top.

10% gels were run for 3 hours at 1000V and 20mA, 8% gels for 24 hours at 2500V and 20mA and 8% 'wedge' gels for 5 hours at 1500V and 30mA, in 1x, 2x and 1xTris-borate buffer, respectively.

After the samples had run for the required length of

* Final conditions, constant power.

time, the gel frame was removed, and the sequencing gel transferred to a backing sheet of used X-ray film. New X-ray film was then exposed to the gel for 12-24 hours at -70°C.

By a combination of gel runs, up to 400 bases could be read from one set of reactions.

2.7 Electroelution of DNA fragments

DNA fragments for labelling or cloning experiments were isolated from agarose or polyacrylamide by electroelution. Excised agarose slabs were placed in dialysis tubing containing 0.5xTris-borate buffer and the tubing placed in an electroelution chamber containing 150ml of 0.5xTris-borate buffer. Electroelution was carried out at a current of 20mA for 2 hours. The eluate was extracted with phenol:chloroform and the DNA precipitated in 0.3M NaAc and 2.5 volumes of ethanol at -70°C.

2.8 Southern Blot transfer and hybridisation

Agarose gels containing separated DNAs to be transferred were bathed twice in Blot 1 solution (0.5M NaOH, 0.5M NaCl) followed by two bathings in Blot 2 solution (1.5M NaCl, 0.5M Tris-HCl, pH7.4), each for one hour. The gel was then placed on six layers of 3MM paper, that had been presoaked in

20xSSC. A sheet of 0.45um nitrocellulose of equal dimensions, presoaked in 2xSSC, was laid on top, followed by two layers of 3MM paper. Paper towelling provided a tower of absorbent, and DNA was allowed to transfer overnight. The filter was then briefly soaked in 2xSSC, prior to baking at 80°C in a vacuum oven.

Prior to hybridisation, filters were placed in modified 10xDenhardt's solution (0.2% polyvinyl pyrrolidone, 0.2% Ficoll, 0.2% bovine serum albumin in 3xSSC) together with denatured herring sperm DNA at 0.2mg/ml, and incubated at 65°C for 16-18 hours.

Hybridisation was carried out in 1xDenhardt's solution containing heat denatured, radioactively labelled DNA, and allowed to proceed for 3-4 hours at 65°C. Filters were then washed of unhybridised material by placing in two lots of 2 litres of 2xSSC at 65°C (unless stated otherwise in the text). After drying, X-ray film was exposed to the filters for 1-7 days at -70°C.

2.9 Sequencing reactions

3ul of single-stranded recombinant M13 DNA (template DNA) and 1ul of 1uM 17-mer primer DNA* were made up to 5ul in reverse transcriptase (RT) buffer (0.5M Tris-HCl, pH8.3, 0.2M KCl, 70mM $MgCl_2$, 1mM EDTA and 10mM DTT) and placed at 65°C for 2 minutes. Template and primer were then allowed to anneal at 37°C for 20-30 minutes. The DNA was

* Prepared by J. Tellom at the Centre for Recombinant DNA Research, RSBS

made up to 21ul by the addition of 16ul of RT buffer and 1ul of DNA polymerase I (Klenow). 5ul aliquots were distributed to four tubes containing 0.25ul of alpha-^{32}P-dATP and 1ul of the appropriate (G, A, T or C) reaction mix. The reaction mixes contain the following ratios of dideoxy-/deoxy-nucleotide triphosphates:

nucleotide triphosphate (uM)	G-mix	A-mix	T-mix	C-mix
dGTP	5	50	50	50
dATP	50	5	50	50
dTTP	50	50	5	50
dCTP	5	5	5	5
ddGTP	250	-	-	-
ddATP	-	250	-	-
ddTTP	-	-	250	-
ddCTP	-	-	-	250

For longer sequencing reactions, another set of reaction mixes was used, containing proportionately less dideoxynucleotide triphosphates and using alpha-^{32}P-dCTP rather than alpha-^{32}P-dATP. This set contains the following ratios of dideoxy-/deoxynucleotide triphosphates:

nucleotide triphosphates (uM)	G-mix	A-mix	T-mix	C-mix
dGTP	5	50	50	50
dATP	50	5	50	50
dTTP	50	50	5	50
dCTP	5	5	5	5
ddGTP	200	-	-	-
ddATP	-	200	-	-
ddTTP	-	-	200	-
ddCTP	-	-	-	200

Polymerisation reactions were allowed to proceed for 10 minutes at room temperature (21°C), before addition of 1ul of 'chase mix' (0.5mM each of dGTP, dATP, dTTP and dCTP) to each reaction tube. After a further 10 minutes at 21°C, 8ul of sequencing dye (90% deionised formamide, 0.05% (w/v) bromophenol blue, 0.05% (w/v) xylene cyanol and 10mM Tris-HCl, pH8.0) were added to each tube, to stop the reaction. Samples were boiled for 2½ minutes immediately prior to loading 2-3ul of each reaction mix onto the sequencing gel. If a second gel was to be run using these reaction mixes, the remainder was stored at -20°C for no longer than 24 hours.

2.10 The M13 system for the subcloning and sequencing of DNA molecules

The dideoxy strand-terminating method of Sanger et al., (1977) was used here to determine the sequence of fragments subcloned into the M13 vectors mp8 and mp9 (Messing and Vieira, 1982). These vectors are derived from the filamentous phage M13 and possess the regulatory region of the E. coli *lac* operon and most of the beta-galactosidase structural gene (Messing et al., 1977). A region of multiple cloning sites has been inserted into the structural gene, but does not affect its function. A functional product is detected by the addition of 5-bromo-4-chloro-3-indolyl-beta-D-galacto-pyranoside (X-gal) in the presence of the *lac* operon inducer, isopropyl-beta-D-thiogalactopyranoside (IPTG). The X-gal is hydrolysed by beta-galactosidase to produce 5-bromo-4-chloro-indigo, an insoluble blue dye (Miller, 1972). Insertion of foreign DNA generally destroys the integrity of the message. In this event no functional beta-galactosidase is produced and plaques generated by recombinant M13 phage are therefore white or clear, while those generated by the intact vector are blue. Vector DNA that has been digested with restriction enzymes, prior to ligation of foreign DNA, usually generates a proportion of white plaques (due probably to small deletions of DNA at the cloning site). Control ligations with no ligase or with no insert DNA were routinely performed to determine the efficiency of the reactions. Recombinants containing GOE sequences were identified by probing with insert DNA from the plasmid subclones, p316-8AA or pGOE5.

Chapter 3

ISOLATION AND STRUCTURAL ANALYSIS OF LAMBDA CLONES CONTAINING GOE SEQUENCES.

The isolation of copies of the GOE sequence from the *Drosophila* genome requires the use of a pure, highly specific DNA as a probe. Two points indicate that the Bkm satellite fraction that was used by Singh et al. (1980b) would not be the most appropriate probe. First, since Bkm DNA was extracted from an analytical gradient, it is not necessarily a homogeneous source of GOE DNA, as it is likely to contain other sequence components. Second, although the snake and *Drosophila* GOE sequences are sufficiently similar to cross-hybridise in standard hybridisation conditions (3xSSC at 60-62°C for 3 hours), *Drosophila* GOE sequences are likely to be more similar to each other than to GOE sequences from another species. Therefore, a *Drosophila* GOE sequence is probably more appropriate to use as a probe for isolating further GOE sequences from the *Drosophila* genome than is Bkm DNA.

Five *Drosophila* lambda clones had been isolated by D. Finnegan* from a *Drosophila melanogaster* genomic library (Maniatis et al., 1978) using Bkm DNA. These were designated lambda clones 314, 315, 316, 317 and 319, respectively by Singh et al. (1980b). A preliminary characterisation of these lambda clones by the same workers demonstrated which

* Edinburgh University, UK

restriction fragments contained sequences similar to Bkm DNA. Plasmid subclones of Bkm-containing restriction fragments from any of these lambda clones could therefore be used as hybridisation probes for withdrawing further GOE lambda clones from the genomic library.

3.1 Subcloning of restriction fragments from clones 315 and 316

A 2.45kb Bam H1 restriction fragment from clone 316 and a 10.0kb Bam H1 restriction fragment from clone 315 were each subcloned into the the Bam H1 site of the plasmid vector, pBR322 and the recombinants designated p316-8A and p315-P8*, respectively. Digestion of p316-8A with Eco R1 and and Bam H1 enzymes generated two insert restriction fragments, of sizes 1.8kb and 0.65kb. When the separated restriction fragments were transferred to nitrocellulose (Southern-blotted) and probed with the insert DNA from p315-P8, only the 1.8kb restriction fragment hybridised, showing that this contained the GOE sequence. This restriction fragment was further subcloned into the Eco R1 and Bam H1 sites of the vector pBR328, and the recombinant designated p316-8AA. At this stage, p316-8AA was the smallest of the plasmid recombinants known to contain a GOE sequence and was therefore used to probe the *Drosophila* genomic library.

* For a summary of the designations and derivations of all recombinant plasmids, see table 3.3

3.2 Isolation of lambda clones from the *Drosophila melanogaster* (Canton S) embryonic DNA library.

About 20,000 lambda recombinants from the *Drosophila* embryonic DNA library (Maniatis et al., 1978) were screened with p316-8AA DNA that was radio-actively labelled using alpha-^{32}P-dCTP and the random primer method of Whitfeld et al. (1982). 29 initial positives were identified, inoculated into 1ml of SM medium and plaque-purified by further hybridisation with p316-8AA DNA. From these plaque-purified positives, nine stocks were established and the clones designated lambda 40-48. (The remaining 20 positives gave weak signals and were not pursued further). In addition to these, seven lambda clones, identified on the basis of hybridisation with p316-8A DNA, were isolated by G. Miklos in a separate screening experiment.

3.3 Determination of restriction enzyme maps for the GOE-containing lambda clones

3.3.1 Restriction maps of plasmid subclones containing GOE sequences.

Clones that had been obtained from three separate library screens were investigated. The first screen had used Bkm DNA

as a probe (Singh et al., 1980b), the second p316-8AA DNA and the third p316-8A DNA. DNA was prepared from 20 of the clones and these are listed in the table below:

Isolated using:	Bkm DNA	p316-8AA DNA	p316-8A DNA
Clone no.	314	40	1
	315	41	7
	316	42	17
	317	43	18
	319	44	28
		46	32
		47	103
		48	

The DNAs of some of these clones were probed with p316-8AA to confirm that each contained a GOE sequence (figure 3.1).

In order to determine how many of these lambda clones share the same GOE sequence, restriction enzyme maps of plasmid subclones containing GOE were first constructed.

The Eco R1 or Bam HI restriction fragments containing the GOE sequence for a number of the lambda clones were subcloned separately into the plasmid vector, pBR322. The resulting recombinants are listed overleaf.

Source clone	Insert size (kb)	Plasmid name
lambda 314	3.90	p314-4
lambda 315	5.10	p315-11
lambda 316	2.45	p316-8A
lambda 319	11.00	p319-13
lambda 47	3.40	p47-18
lambda 48	2.75	p48-13
lambda 28	8.90	p28A

To determine restriction enzyme maps, the plasmids were digested with enzymes Eco Rl, Bam Hl, Hind III and Pst 1. If a plasmid had few if any sites for these enzymes then Sal 1, Xba 1 and Xho 1 were also used. Except for Xba 1 and Xho 1, which are absent, all these enzymes sites are present once only in the vector.

Restriction maps were constructed on the basis of the sizes of restriction fragments produced on digestion with these enzymes, both singly and in combination. Restriction digests were also Southern-blotted, and probed with the insert DNA from p316-8AA, to identify those fragments carrying a GOE sequence. p315-11 and p47-18 did not yield appropriate restriction enzyme fragments flanking the GOE sequence, and instead were used whole to probe the lambda clones.

Restriction maps are shown in figure 3.2 and the restriction fragments used to probe the lambda clones indicated.

3.3.2 Identification of overlapping lambda clones.

Lambda clones were digested with Eco Rl, the restriction fragments separated on agarose gels and Southern-blotted. They were probed with the restriction fragments that flank the GOE sequences and that are indicated in figure 3.2. The results of the hybridisation experiments are presented in figures 3.3 and 3.4 and summarised in Table 3.1. (Some lambda clones were not included in these experiments because their restriction enzyme digestion patterns showed them to be duplicates of other clones. For example, clone 43 is equivalent to clone 47 and clones 7 and 42 are equivalent to 316).

On the basis of these experiments the collection of 20 clones can be arranged into six sets, which are listed below. The clones within a set share sequences that are absent from the other clones.

Set	Clones	Isolated with Bkm	Isolated with p316
GOE4	314,317,40 46,32,103	314,317	40,46,32,103
GOE5	315,17	315	17
GOE6	316,41,42, 48,7,18	316 7,18	41,42,48,
GOE7	43,47		43,47
GOE8	28		28
GOE9	319,44	319	44

Table 3.1. Summary of cross-hybridisation results for sequences flanking the GOE regions

	Lambda clones															
Probe	314	317	1	32	103	316	41	48	18	319	44	315	17	47	28	
p314-4A [A]	+++	+++	+++	+++	+++	-	-	-	-	-	-	-	-	-	-	
p314-2 [B]	+++	+++	+++	+++	+++	-	-	-	-	-	-	-	-	-	-	
p316-8A (Xho-Bam fragment). [C]	-	-	-	-	-	+++	+++	+++	+++	-	-	-	-	-	-	
p48-11 [D]	-	-	-	-	-	+++	+++	+++	+++	-	-	-	-	-	-	
p319-1 [E]	-	-	-	-	-	-	-	-	-	+++	+++	-	-	-	-	
p319-18 (Sac-Pst fragment). [F]	-	-	-	-	-	-	-	-	-	+++	+++	-	-	-	-	
p315-T22	-	-	-	-	-	-	-	-	-	-	-	+++	+++	-	-	
p47-18	-	-	-	-	-	-	-	-	-	-	-	-	-	+++	-	
p28A	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+++	

+++ strong hybridisation.
+ weak hybridisation.
- no detectable hybridisation.

3.3.3 Restriction maps of the lambda clones.

Restriction enzyme maps were determined for some of the lambda clones from each set, using the same set of restriction enzymes as were used for determining the plasmid maps. To detect any small Eco R1 fragments that were not visible under UV light, restriction enzyme fragments were end-labelled with alpha-^{32}P-dATP, and separated on agarose and polyacrylamide gels. The gels were dried and used directly to expose autoradiograph film (figure 3.5). The sizes of restriction fragments were determined with respect to the standard size markers of wild-type lambda digested with Hind III and pBR322 digested with Hinf I, and are listed in table 3.2. Ambiguities in the restriction map were resolved by probing Southern-blotted digests with appropriate plasmid subclones.

The restriction maps are presented in figure 3.6. It is apparent that the clones within a set have similar restriction maps and are therefore overlapping. This confirms that the six sets of clones represent six separate GOE sequences.

Eco RI and other enzyme restriction fragments that together represent all of the DNA that includes and flanks the different GOE copies were subcloned into the vectors, pBR322 or pBR328. Their designations are summarised in table 3.3 and their locations relative to the lambda clone restriction maps shown in figure 3.7.

Table 3.2. Eco RI restriction fragments of lambda clones

Clone no.	314	317	315	17	316	48	41	44	319
Restriction fragment size (kb)	5.65	10.60	8.60	5.60	3.90	5.40	7.40	8.00	11.00
	4.25	4.25	5.10	5.10	2.55	3.90	5.40	2.75	1.76
	3.90*	0.70	0.50	3.80	2.45	2.75	2.50	2.50	0.24
		0.12	0.30	0.50	2.45	0.70	2.35	1.95	
					1.40	0.52		0.95	
					0.70	0.11		0.44	
								0.33	

Clone no.	47	18	28	32	103a	1
Restriction fragment size (kb)	3.50	5.40	8.90	4.25	8.00	8.00
	3.40	4.40	4.70	4.25	4.50	4.25
	3.25	2.70	0.51	4.10	4.25	2.65
	1.42	1.74	0.43		1.44	
	0.80	0.82				
	0.63					
	0.42					
	0.27					
	0.22					

* Underlined restriction fragments hybridise to p316-8AA or p315-T22 DNA.

3.4 Estimation of the copy number of GOE sequences in *Drosophila*

Have all GOE copies present in the library been collected? Except for GOE8, all sets are represented by two or more lambda clones. This suggests that most of the GOE copies present in the *Drosophila* library have in fact been isolated. One would conclude from this that there are not many more than 6 different copies of the GOE sequence in the *Drosophila melanogster* genome.

However, when *D. melanogaster* genomic DNA is digested with restriction enzymes (e.g Alu I) and probed with GOE sequences (either Bkm DNA or p316-8AA DNA) there is strong hybridisation to high molecular weight fragments (e.g. Singh et al., 1980b and figure 3.8). The intensity of hybridisation is far in excess of what one would expect if only six dispersed copies were present. This would suggest that the library does not contain a full complement of genomic sequences. A fraction of the genome, which incorporates much of the high molecular weight sequences responsible for hybridisation to Bkm or GOE DNA, has probably been excluded.

Finnegan et al. (1977) estimate that much of the highly repeated DNA sequences, which make up 20% of the *D. melanogaster* genome, is not represented in this library. As the GOE sequences in the high molecular weight fraction must be very numerous to produce such intense hybridisation, they also are probably excluded from the library along with the

other highly repeated sequences. This could explain why so few copies of the GOE sequence have been isolated even though the genome probably contains thousands of copies.

If GOE DNA is predominantly composed of poly(GATA) sequences (Epplen et al., 1983a,b and Singh et al., 1984), it may be capable of forming stable duplexes with some of the *Drosophila* satellite sequences. This could explain the relatively intense hybridisation to high molecular weight restriction fragments.

Four major *D. melanogaster* satellite DNAs (satellites I, II, III and IV) have been characterised (Brutlag, 1980). Satellite III is a tandem repeat of a 359bp sequences (Hsieh and Brutlag, 1979) that contains only one GATA tetra-nucleotide. Unless GOE DNA is composed of other sequences in addition to poly(GATA), it would not hybridise to satellite III DNA under the conditions used here.

Satellites I, II and IV are tandem repeats of pentameric and septameric sequences. These are listed below and it can be seen that none of them contain a GATA (or TATC) tetra-nucleotide.

Satellite I	1.672g/cm^3	5' AATAT 3'	(Brutlag and
		5' AATATAT 3'	Peacock, 1979).
Satellite II	1.686g/cm^3	5' AATAACATAG 3'	(Endow et al., 1975).
Satellite IV	1.705g/cm^3	5' AAGAG 3'	(Fry and
		5' AAGAGAG 3'	Brutlag, 1979).

However, if one allows for a G residue to pair with a T,

then the GATA strand of a GOE sequence could form a duplex molecule with one of the satellite I strands, namely:

GATA-GATA
$TTAT_ATTAT_A$

Such a hybrid molecule may not be vary stable (since there is an unpaired residue at every fifth base), but the hybridisation conditions used here could be sufficient to produce the observed signal.

Though the 'GATA' strand is potentially capable of hybridising to satellite I DNA, the 'TATC' strand is unlikely to be. The possible hybrids between the 'TATC' strand and either strand of the satellite I sequence can only be paired at 3 out of every 5 bases, i.e.

TAT^C-TAT^C
$ATA_{TT}ATA_{TT}$

TAT^C-TAT^C
$T^{ATA}{}_{AT}{}^{ATA}{}_A$

If satellite I were responsible for the hybridisation of GOE DNA to the high molecular weight restriction fragments, one would expect much stronger hybridisation with the 'GATA' strand than with the 'TATC' strand.

This was tested by using single-stranded DNAs to probe Canton S genomic DNA. The probes were manufactured from M13 recombinants that contained the GOE6 sequence in both orientations (see Chapter 4). The single stranded recombinant containing the 'GATA' strand can therefore be used to generate

a radioactively-labelled poly(TATC) DNA molecule (and vice versa for the recombinant containing the 'TATC' strand). The results of these hybridisation experiments are presented in figure 3.9 and show that both 'GATA' and 'TATC' strands produce similar hybridisation patterns of equal intensity. Therefore, satellite I DNA cannot be responsible for the hybridisation of GOE sequences to high molecular weight restriction fragments. Possibly there are other 'cryptic' satellite, or satellite-like, sequences which may be sufficiently homologous to poly(GATA) to hybridise. Of course, it is also likely that the high molecular weight hybridising material *is* poly(GATA) DNA.

At first, this might seem to indicate that the *Drosophila* GOE family of repeated sequences is not a good system for a study of repeated sequence function, since the copy number is so high. However, this objection can be obviated by the following arguments.

First, the bulk of the hybridising material is confined to high molecular weight restriction fragments. These fragments must contain an abnormal distribution of nucleotides on a large scale. Such sequences are unlikely to contain coding or genic sequences.

Second, this genomic library has been used successfully to isolate numerous lambda clones of euchromatic and genic sequences, both dispersed and contiguous. Therefore, that fraction of the GOE sequence complement that is euchromatic in

location and interspersed with genic sequences should also be extractable.

Third, evidence will be presented later (Chapter 6) to show that the highly repeated, high molecular weight fraction of the GOE sequence family is absent in other *Drosophila* species, while that fraction of GOE sequences residing in discrete, lower molecular weight restriction fragments is maintained. This latter class of GOE sequences is more likely to have a function that is shared by other *Drosophila* species and perhaps by other organisms as well.

3.5 Are any of the collected GOE sequences from region 19F-20AB?

Singh et al. (1980b) reported that Bkm DNA hybridised *in situ* solely to one region of *D. melanogaster* salivary gland polytene chromosomes. This was region 19F-20AB on the X chromosome. On the other hand, as well as hybridising to this region, the *Drosophila* GOE-containing lambda clones hybridised additionally to one of four other euchromatic sites. Namely:

Clone 314	hybridises	to bands	19F-20AB	*and*	38B (chromosome 2)
Clone 317	"	"	"		38B (chromosome 2)
Clone 315	"	"	"		95A (chromosome 3)
Clone 316	"	"	"		11E (chromosome 1)
Clone 319	"	"	"		52F (chromosome 2)

A cloned 200bp *Drosophila* GOE sequence was used to probe

Drosophila polytene chromosomes and it was found that this also hybridised only to region 19F-20AB (G. Levinson, pers. comm.). Why do lambda clones that contain GOE sequences hybridise to sites that GOE sequences themselves will not? One possible explanation for this (favoured by Jones and Singh, 1982) is that some rearrangment of the regions around the GOE sequences has occurred during development, such that the flanking sequences were transferred to the other hybridisation sites, while the GOE sequences themselves remained at region 19F-20AB.

An alternative and simpler explanation is that these lambda clones do originate from these other locations, but the GOE sequences that they contain are either too small or heterologous to be detected by either the Bkm or GOE DNAs *in situ*, but not in a library screen. For example, a unique restriction fragment from lambda clone 316 hybridises *in situ* to band 11E only and not to region 19F-20AB (G. Miklos, pers. comm.). This shows that lambda clone 316 is derived from 11E and contains a GOE sequence that is not detectable *in situ*. Obviously it is possible that the other lambda clones also are not from 19F-20AB, and this possibility needed to be tested.

To determine whether any of the other GOEs (including those that have not been localised cytogenetically) are not derived from region 19F-20AB, non-GOE-containing subclones were probed to Eco RI digests of genomic DNAs of male and female flies (extracted from adult heads, which, being predominantly composed of neural tissue, are diploid).

Fragments located at region 19F-20AB, and therefore on the X chromosome, will give twice as strong a signal to female DNA than to male DNA, because twice as much X-chromosomal material is present. To calibrate the male and female DNA amounts, the DNAs were also probed with sAC1 DNA (Goldberg et al., 1980) - a plasmid subclone of a 4.7kb Eco RI fragment containing the *Adh* gene which is both unique and autosomal. Figure 3.10 shows that p47-13 contains a unique fragment. Densitometer tracings were made of the sAC1 and p47-13 bands from the male and female DNAs, and the areas of the four peaks estimated. If p47-13 is autosomal then the ratio of the area between female and male DNAs should be the same as for sAC1. If p47-13 is from the X chromosome, then the female to male ratio should be twice that of sAC1. In fact, the ratio between p47-13 and sAC1 is less than one and one must conclude that p47-13 is autosomally derived. This means that GOE7 also is an autosomal copy of the GOE sequence. Similar experiments using p314-2 and p315-11B DNAs showed that the GOE4 and GOE5 copies are also autosomally derived and probably reside at bands 38B and 95A, respectively.

Similar experiments could not be performed for the lambda 28 clone because the plasmid tested (p28B) hybridises strongly to three restriction fragments and more weakly to a large number of others. Since p28B is derived from a terminal restriction fragment of the lambda 28 clone, one cannot determine which of the hybridising genomic bands corresponds to it. However, even if GOE8 does reside in the 19F-20AB

region, it cannot be the only copy. The hybridisation of p316-8AA to all the lambda GOE copies is fairly consistent, whereas it is very intense *in situ* to the 19F-20AB region. More than one equivalent copy of GOE must be located there.

3.6 Isolation from a *D. melanogaster* (FF strain) genomic library of clones containing GOE sequences.

The sequencing of the GOE copies collected will provide a measure of the structural variation of this family of repetitive sequences within a genome. This variation would need to be taken into account if some function is to be attributed to the family as a whole. It can be argued, though, that only one or two copies need be functional, so that selective constraints would allow them to diverge less than their non-functional relatives. For example, families of genes are composed of both transcribed structural genes and non-transcribed pseudogenes. This objection could be resolved by comparing the sequences of GOE elements between genomes. Ideally, all the GOE copies from two genomes should be compared but, as an initial approach, it was decided to ensure that the copy equivalent to GOE6 at least should be isolated from the genome of a wild strain of fly. Strain FF was derived from a single female collected in New South Wales for other purposes (Lewis and Gibson, 1978). It carries the fast (Adh^F) allele of the *alcohol dehydrogenase* locus. Preliminary

sequence data had shown GOE6 to contain the longest unbroken stretch of GATAs as compared to the published sequences. It would be most parsimonious then to derive these other sequences from GOE6, rather than *vice versa*, and so GOE6 is more likely to represent an ancestral copy.

To isolate a particular repeated sequence from a genome, an adjacent, unique sequence is required as a probe. Plasmid p48-11, which contains a 0.7kb Eco RI-Bam HI restriction fragment directly adjacent to the 1.8kb restriction fragment containing GOE6, is unique in both Canton S and FF genomes (see section 6.1). This recombinant was therefore used to probe a library of FF genomic sequences.

The library was prepared by digesting FF embryonic DNA with Eco RI and ligating the restriction fragments to the corresponding sites of the vector, lambda gt10 (C. Collet, pers. comm.). The vector is able to accept restriction fragments in the range 2.5-7.0kb only, and so the library cannot contain a full complement of genomic DNA. As p48-11 resides in a 2.7kb restriction fragment in the FF genome, it should however be present in the library.

Approximately 10,000 plaques were transferred to duplicate sets of nitrocellulose filters. One set was probed with p48-11 and the other with p316-8AA to detect other GOEs apart from GOE6. Two different lambda clones were isolated and purified. They were designated lambda FF3 and FF4, respectively.

DNA from these lambda clones was digested with Eco RI to liberate the *Drosophila* restriction fragments. Lambda FF3 contained a single 2.7kb insert restriction fragment, whereas lambda FF4 contained three Eco RI restriction fragments of sizes 4.2, 1.75 and 0.60kb, respectively. Probing these restriction fragments with pGOE5 showed that the 2.7kb restriction fragment from FF3 and the 4.2kb from FF4 contain GOE sequences. For further analysis, these two Eco RI restriction fragments were subsequently subcloned into the Eco RI site of pBR322 and the two resulting recombinants designated pFF1 (from lambda FF3) and pFF12 (from lambda FF4).

pFF1 and pFF12 each yielded two *Drosophila* restriction fragments on digestion with Eco RI and Bam HI enzymes. Only the 0.8kb restriction fragment from pFF1 hybridised to p48-11. The other (1.9kb) restriction fragment and a 0.9kb restriction fragment from pFF12 hybridised with p316-8AA. pFF1 is therefore the FF equivalent of the Canton S subclone, p48-13, and must contain the GOE6 equivalent. The two *Drosophila* Eco RI-Bam HI restriction fragments from pFF1 were subcloned into pBR322, and designated pFF2 (0.8kb) and pFF3 (1.9kb).

To see whether pFF12 corresponded to any of the other Canton S GOEs that have been isolated, the non-GOE (3.3kb) Eco RI-Bam HI restriction fragment was probed to the six plasmids containing the Canton S GOE sequences (namely, p314-4, p315-11, p48-13, p47-18, p28A and p319-18), but no hybridisation was detected to any of these plasmids. Two possibilities can

account for this:

i) pFF12 contains a GOE sequence that is derived from another part of the genome and is consequently surrounded by DNA unrelated to that surrounding the other GOE sequences.

ii) the GOE sequence in pFF12 is equivalent to one of those already isolated from the Canton S genome, but in the FF genome it has transposed to another region.

These two possibilities can only be resolved by sequencing the GOE sequence in pFF12. A restriction map was constructed for this plasmid, using the strategy described earlier in this chapter and the location of the GOE sequence was identified by probing with the insert from p316-8AA. This map is presented in figure 3.12 and shows the GOE sequence to lie in a 0.7kb Eco RI-Pst I restriction fragment.

3.7 Summary

An intensive screen of a *D. melanogaster* (Canton S) genomic library has revealed six distinct copies of the GOE sequence. Two GOE copies have also been isolated from a partial library of a wild strain (FF) of *D. melanogaster*. Their nucleotide sequences can now be determined and an estimate made of the variation between the copies within a genome.

Table 3.3. Plasmid designations and derivations

CLASS 4

Clone	Vector	Insert size (kb)	Insert ends	Source
p314-2	pBR322	4.25	E-E	314
p314-4	pBR322	3.90	E-E	314
p314-4A	pBR328	2.60	E-P	p314-4
p314-4B	pBR328	1.30	E-P	p314-4
p314-8	pBR322	5.60	E-E	314
p317-12	pBR322	10.60	E-E	317
p317-1	pBR322	5.50	E-P	p317-12

CLASS 5

Clone	Vector	Insert size (kb)	Insert ends	Source
p315-8	pBR322	8.10	E-E	315
p315-11	pBR322	5.10	E-E	315
p315-11B	pBR322	2.00	E-B	p315-11
p315-T22	pBR322	3.10	E-B	p315-11
p315-P8	pBR322	10.00	B-B	315
pGOE5	pBR322	0.45	H-H	p315-T22
p17B	pBR322	3.80	E-E	17
p17C	pBR322	0.50	E-E	17

CLASS 6

Clone	Vector	Insert size (kb)	Insert ends	Source
p316-8A	pBR322	2.40	B-B	316
p316-8AA	pBR328	1.90	E-B	p316-8A
p316B6	pBR322	2.45	E-E	316
p316B9	pBR322	2.55	E-E	316
p316C	pBR322	1.80	E-E	316
p316D	pBR322	0.70	E-E	316
p48-3	pBR322	5.40	E-E	48
p48-4	pBR322	3.90	E-E	48
p48-13	pBR322	2.70	E-E	48
p48-11	pBR322	0.80	E-B	p48-13
p41-11	pBR322	2.50	E-E	41
p41-20	pBR322	7.40	E-E	41
pFF1	pBR322	3.75	E-E	FF3
pFF2	pBR322	0.80	E-B	pFF-1
pFF3	pBR322	1.90	E-B	pFF-1

B = Bam HI, E = Eco RI, H = Hae III, P = Pst I.

Table 3.3 (contd.)

CLASS 7

Clone	Vector	Insert size (kb)	Insert ends	Source
p47-4	pBR322	0.60	E-E	47
p47-12	pBR322	0.80	E-E	47
p47-13a	pBR322	1.40	E-E	47
p47-13	pBR322	3.25	E-E	47
p47-16	pBR322	3.70	E-E	47
p47-18	pBR322	3.40	E-E	47

CLASS 8

Clone	Vector	Insert size (kb)	Insert ends	Source
p28A	pBR322	8.90	E-E	28
p28B	pBR322	4.70	E-E	28
p28-5	pBR322	0.51	E-E	28
p28-12	pUC8	2.20	P-P	p28A

CLASS 9

Clone	Vector	Insert size (kb)	Insert ends	Source
p319-13	pBR322	11.00	E-E	319
p319-T1	pBR322	2.30	B-B	319
p319RP1	pBR322	1.60	E-P	p319-13
p319-5	pBR322	2.70	P-P	p319-13
p319-8	pBR322	1.70	P-P	p319-13
p319-18	pBR322	3.20	P-P	p319-13
p44-4	pBR322	8.0	E-E	44
p44-10	pBR322	1.95	E-E	44
p44-11	pBR322	2.40	E-E	44
p44-12	pBR322	2.75	E-E	44
p44-8	pBR322	4.00	E-E	44

CLASS 12

Clone	Vector	Insert size (kb)	Insert ends	Source
pFF12	pBR322	4.2	E-E	FF4
pFF12-1	pBR322	3.3	E-B	pFF12
pFF12-2	pBR322	0.9	E-B	pFF12

B = Bam HI, E = Eco RI, H = Hae III, P = Pst I.

Figure 3.1

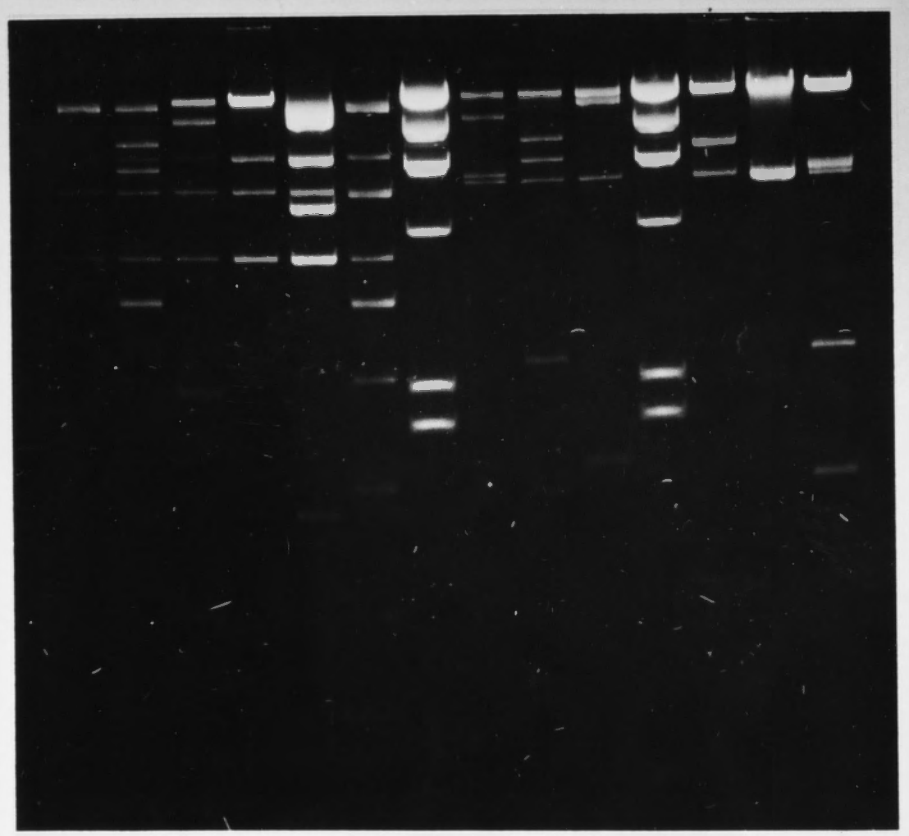
Probing of lambda clones with p316-8AA. 1-2ug of DNA extracted from lambda clones were digested to completion with 2 units of Bam HI or Hind III enzymes and the restriction fragments separated on 1% agarose gels and Southern blotted as described in Materials and Methods.

a) Ethidium bromide staining pattern under UV light.

b) Autoradiograph after hybridisation with p316-8AA DNA (16 hours' exposure).

All six lambda clones contain single fragments that hybridise to p316-8AA DNA.

a.



b.

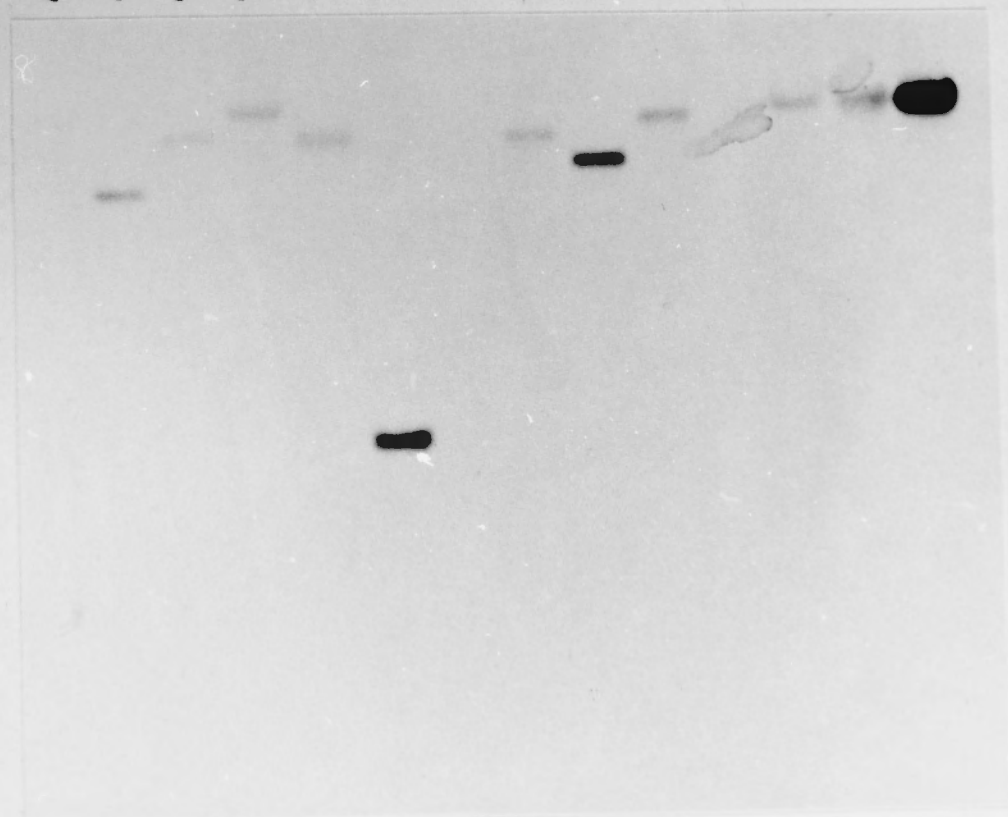


Figure 3.2

Restriction maps of the regions that surround those restriction fragments (heavy lines) known to hybridise to p316-8AA or p315-P8 DNA. Open boxes delineate the fragments (A to F) that were used to probe the lambda clones in figures 3.3 and 3.4.

Restriction fragment A = p314-2

Restriction fragment B = p314-4A

Restriction fragment C = 1.5kb Bam HI-Xho I fragment from p316-8A

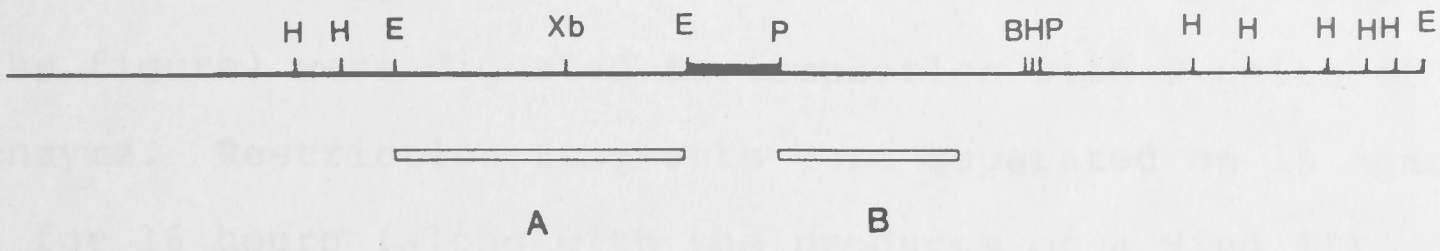
Restriction fragment D = 0.9kb Bam HI-Eco RI fragment from p48-13

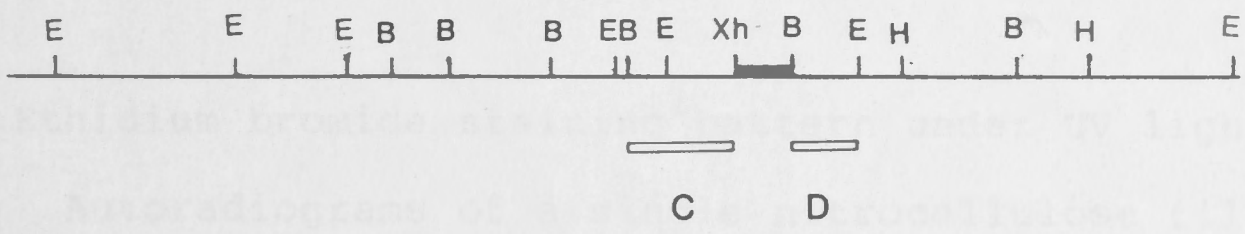
Restriction fragment E = 0.9kb Pst I-Sac I and 0.8kb Sac I fragments from p319-18

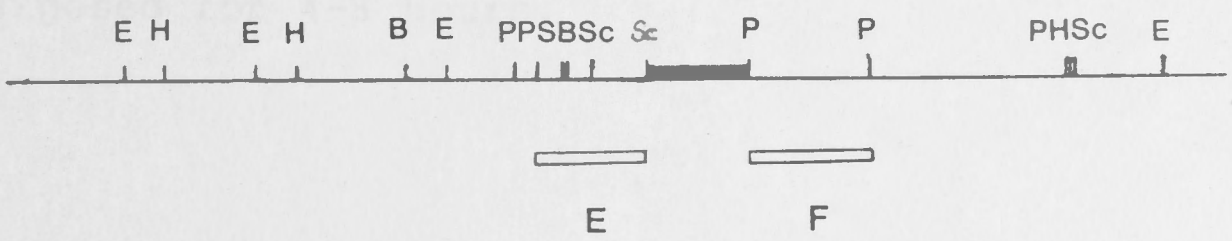
Restriction fragment F = p319-5

Restriction fragments C, D and E were electro-eluted from 1% sea-plaque agarose gels as described in Materials and Methods. The remaining fragments were labelled as part of the original recombinant plasmid(s).

Restriction fragments used to probe the lambda clones







B Bam HI

E Eco RI

H Hind III

P Pst I

Sc Sac I

Xb Xba I

Xh Xho I

1kb

Figure 3.3

Hybridisation to the lambda clones of the restriction fragments that flank the GOE-containing regions. 0.5-1ug of the GOE-containing lambda clones (clone numbers as indicated in the figure) were digested to completion with 2 units of Eco RI enzyme. Restriction fragments were separated on 1% agarose gels for 16 hours (along with the products of a Hind III digestion of lambda C1857 to provide size markers). After denaturation, DNA was transferred to a nitrocellulose filter, as described in Materials and Methods.

a) Ethidium bromide staining pattern under UV light.
b-e) Autoradiograms of a single nitrocellulose filter that was successively probed, washed and reprobed with the following radioactively-labelled DNAs: b) fragment A, c) fragment B, d) fragment C and e) fragment D. Autoradiograms were exposed for 4-8 hours.

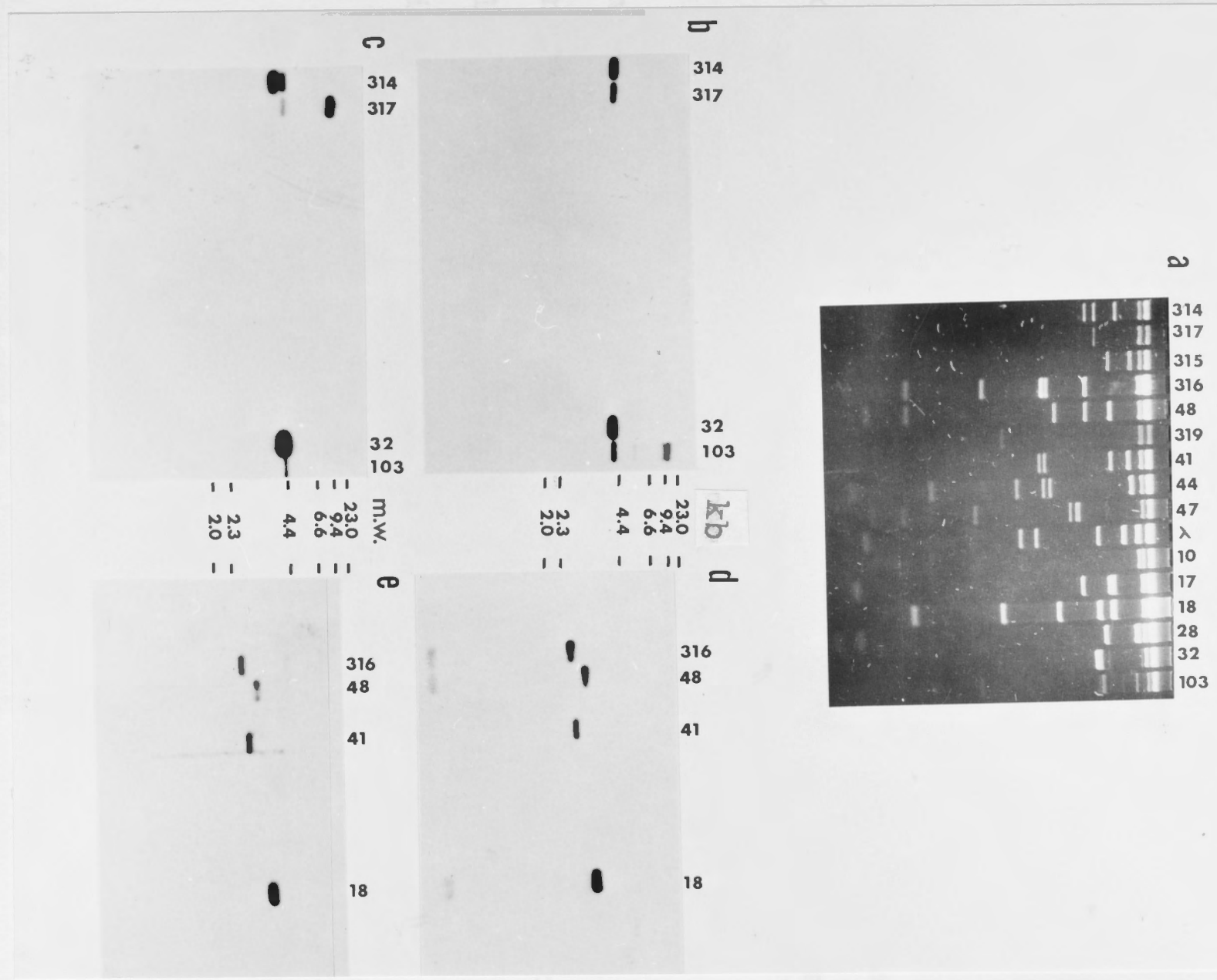

a
314
317
315
316
48
319
41
44
47
λ
10
17
18
28
32
103
b
314
317
32
103
kb
23.0
9.4
6.6
4.4
2.3
2.0
c
314
317
32
103
m.w.
23.0
9.4
6.6
4.4
2.3
2.0
d
316
48
41
18
e
316
48
41
18

Figure 3.4

As in figure 3.3, except that only those lambda clones that did not hybridise to restriction fragments A to D were tested.

a) Ethidium bromide staining pattern under UV light.

b-f) Autoradiograms of two nitrocellulose filters that were successively probed, washed and reprobed with two or three of the following radioactively-labelled DNAs: b) p315-T22, c) p47-18, d) fragment E, e) fragment F and f) p28A.

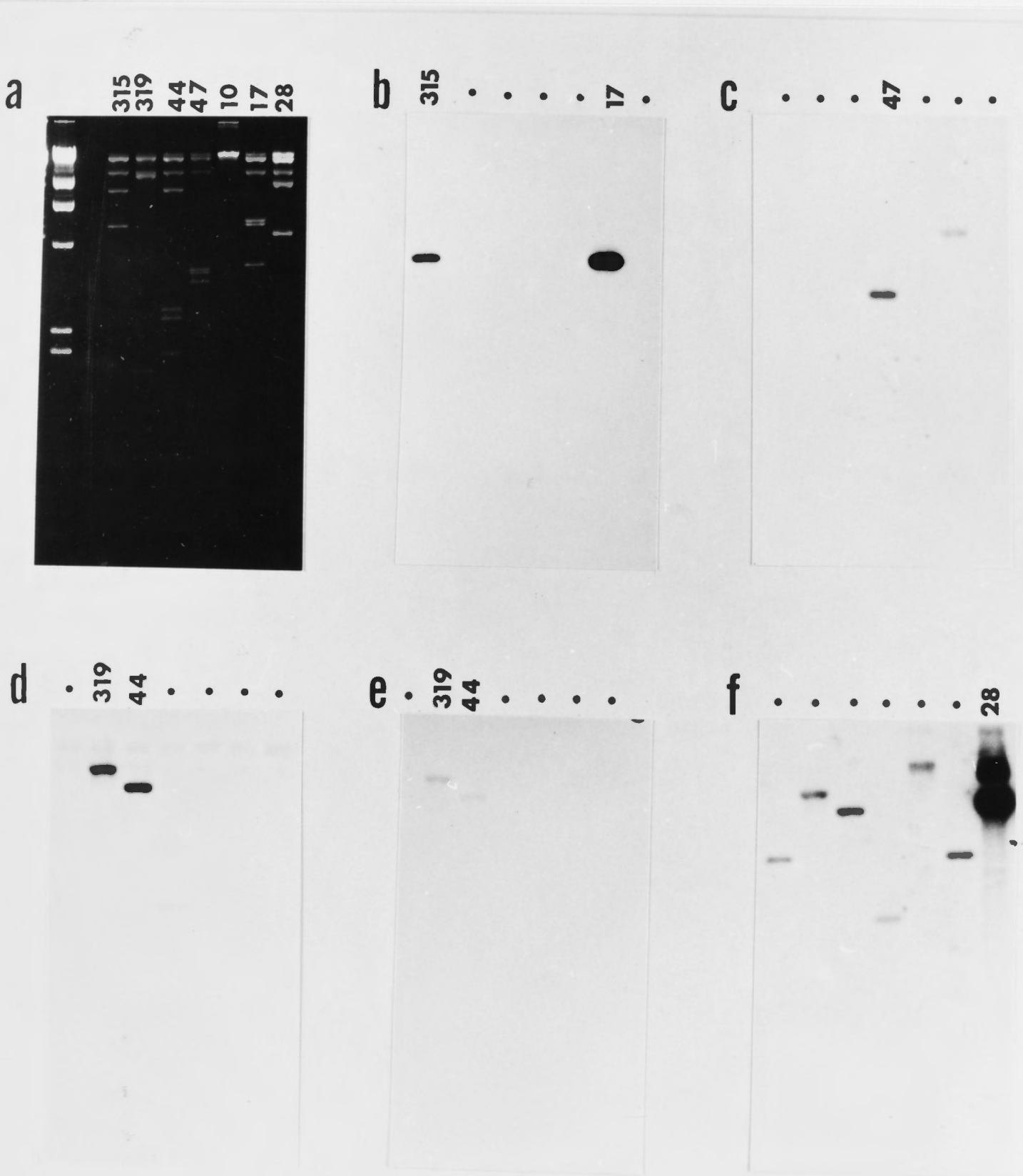

a
315
319
44
47
10
17
28
b
315
17
c
47
d
319
44
e
319
44
f
28

Figure 3.5

Autoradiograms of Eco RI-digested and end-labelled lambda clones. 0.5ug of selected lambda clones (clone numbers as indicated in the figure) were digested with Eco RI and end-labelled as described in Materials and Methods. (Lambda C1857 DNA digested with Hind III and pBR322 DNA digested with Hinf I were similarly treated, to provide size markers).

Restriction fragments were separated on a) 1% agarose and b) 8% acrylamide gels for 16 hours and 3 hours, respectively. Gels were dried under vacuum, and autoradiograms exposed for 4 to 6 hours. Sizes of the marker fragments are indicated to the right of each autoradiogram.

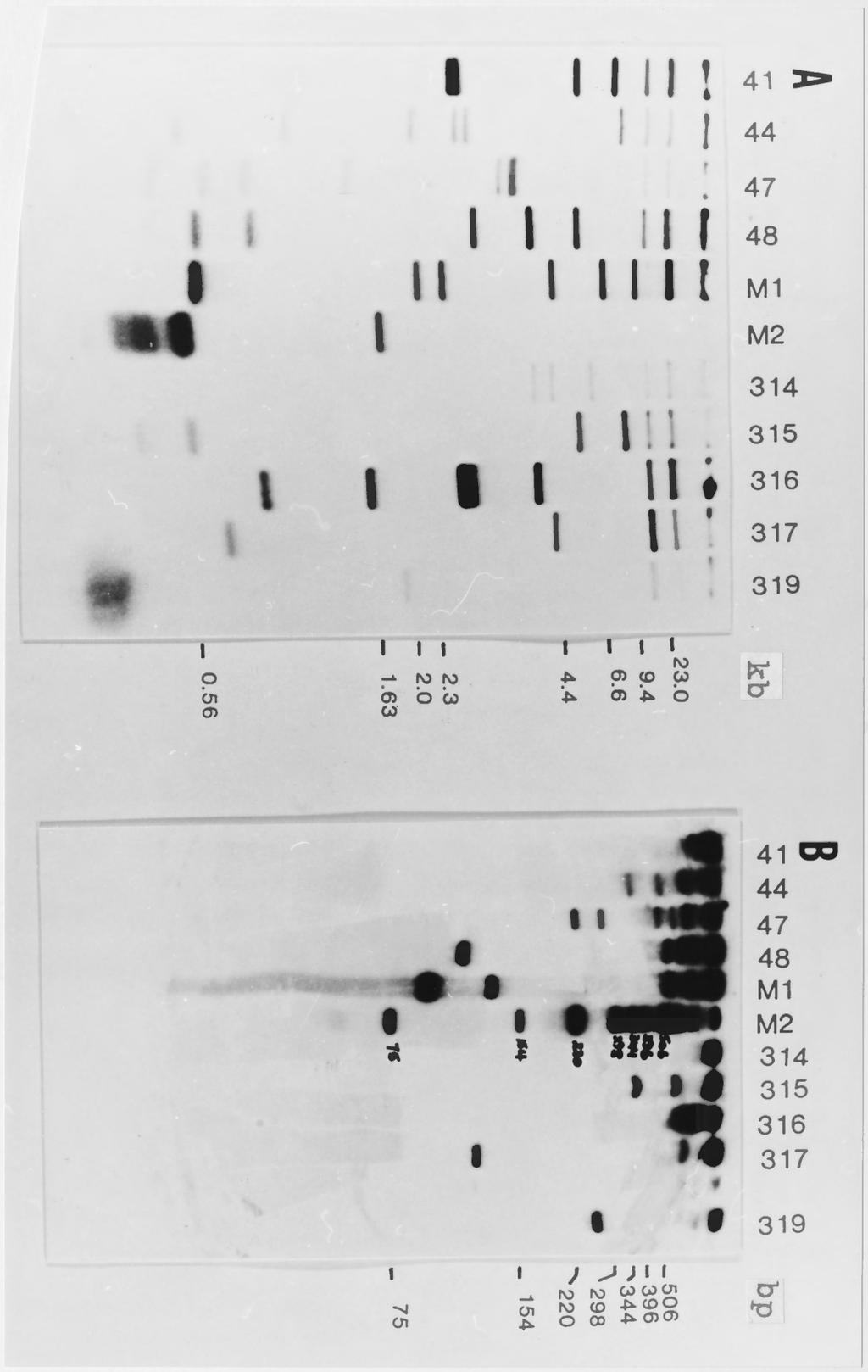
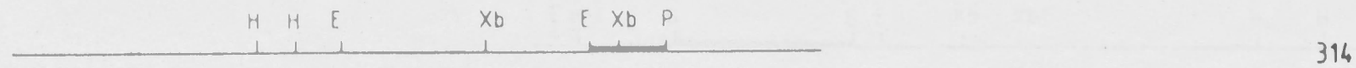

A
41
44
47
48
M1
M2
314
315
316
317
319
kb
23.0
9.4
6.6
4.4
2.3
2.0
1.63
0.56
B
41
44
47
48
M1
M2
314
315
316
317
319
bp
506
396
344
298
220
154
75

Figure 3.6

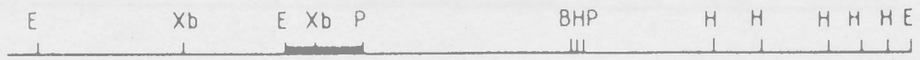
Restriction maps of selected lambda clones from each of the six sets that are described in the text. The keys to the symbols used for each restriction enzyme site are provided in the figures. Heavy lines delineate those regions to which GOE sequences have been localised.

a) GOE4, b) GOE5, c) GOE6, d) GOE7, e) GOE8 and f) GOE9.

A) GOE4



314



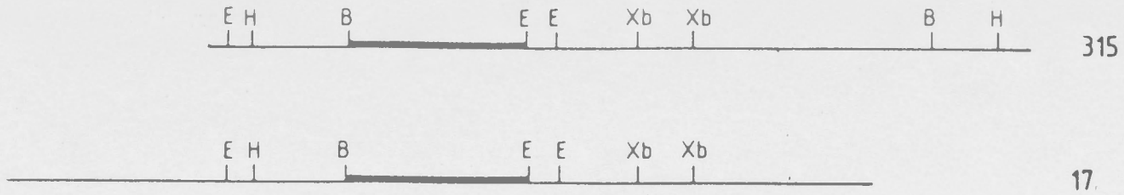
317

1

B BamH1
E EcoR1
H HindIII
P PstI
Xb XbaI

1kb

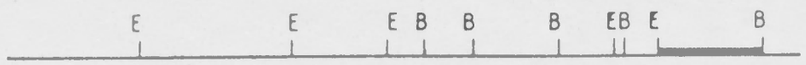
B) GOE5



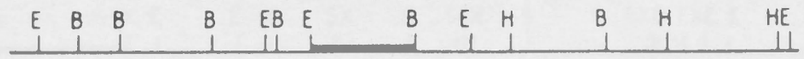
B BamH1
E EcoR1
H HindIII
Xb XbaI

1kb

c) GOE6

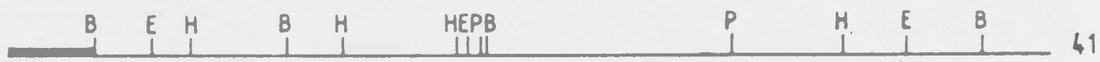


316



48

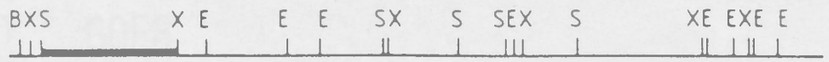
B BamH1
E EcoR1
H HindIII
P Pst



41

1kb

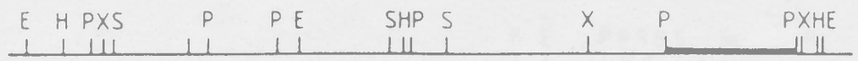
D) GOE7



B BamH1
E EcoR1
P Pst1
S Sal1
X Xho1

1kb

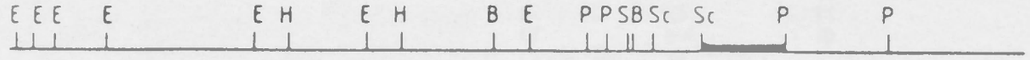
E) GOE8

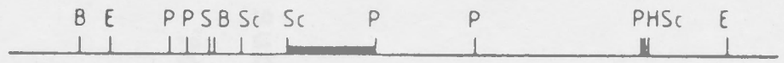


B BamH1
E EcoR1
H HindIII
P PstI
S SalI
X XhoI

1kb

F) GOE9





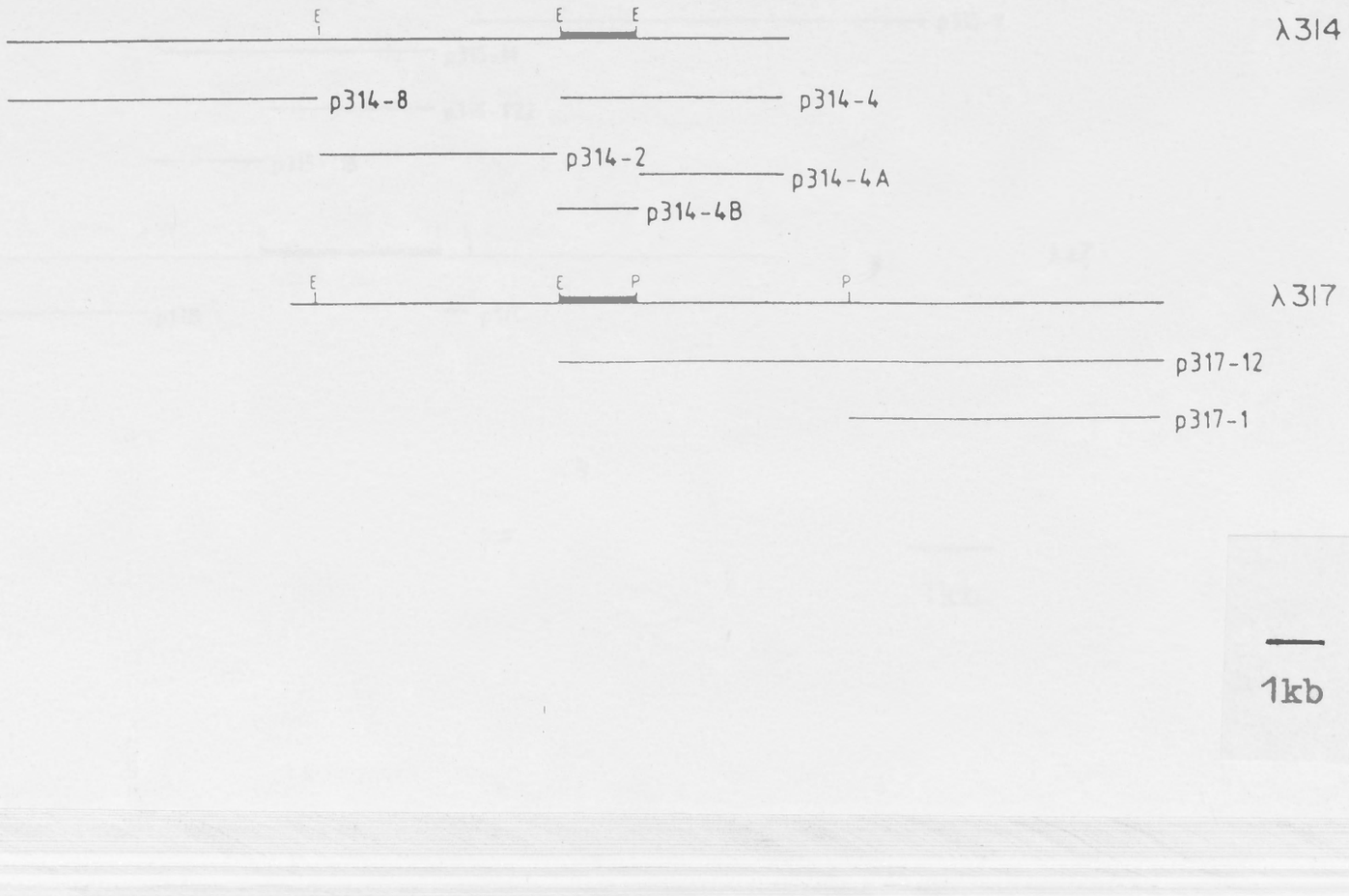
B BamH1
E EcoR1
H HindIII
P PstI
S SalI
Sc SacI

1kb

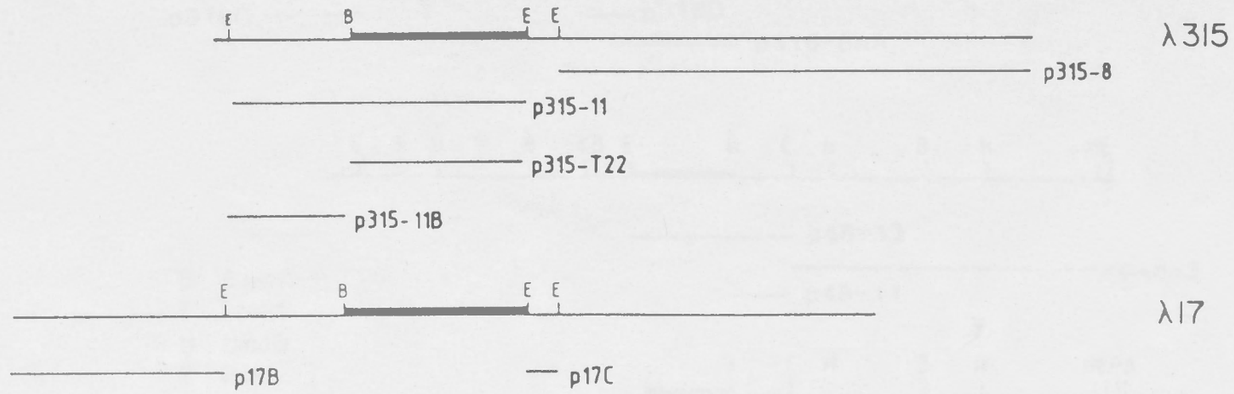
Figure 3.7

Positions of the various plasmid subclones with respect to the lambda clone restriction maps. Table 3.3 gives a complete description of each subclone.

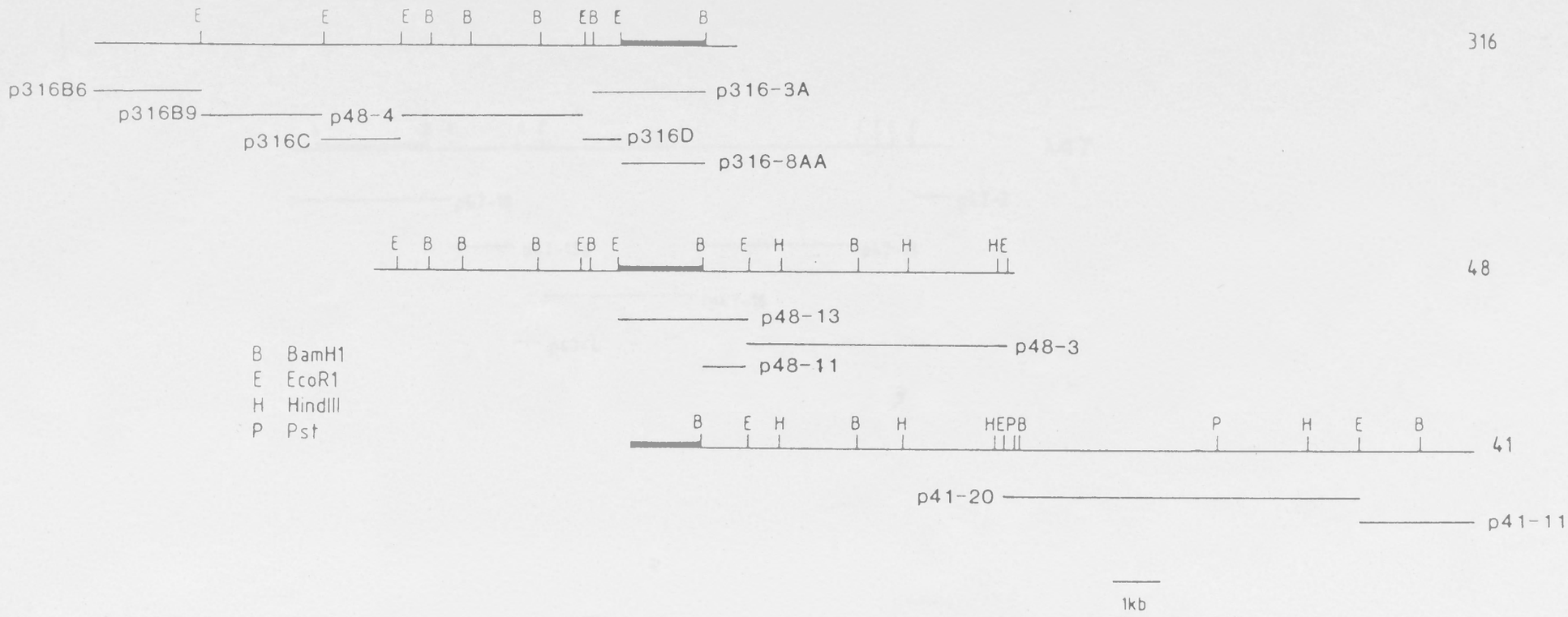
A) GOE4


E
E
E
λ314
p314-8
p314-4
p314-2
p314-4A
p314-4B
E
E
P
P
λ317
p317-12
p317-1
1kb

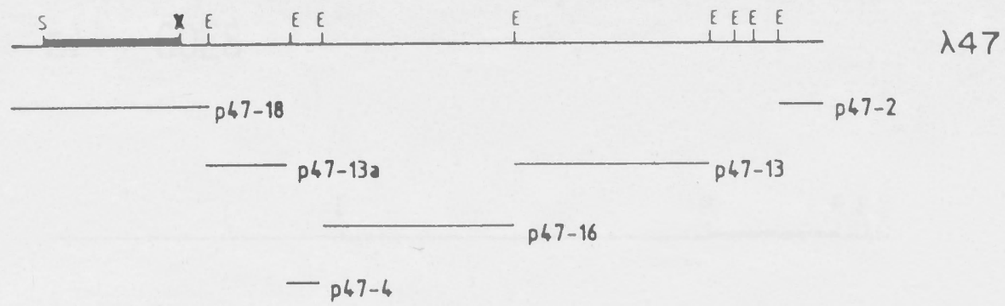
B) GOE5


E
B
E
E
λ315
p315-8
p315-11
p315-T22
p315-11B
E
B
E
E
λ17
p17B
p17C
1kb

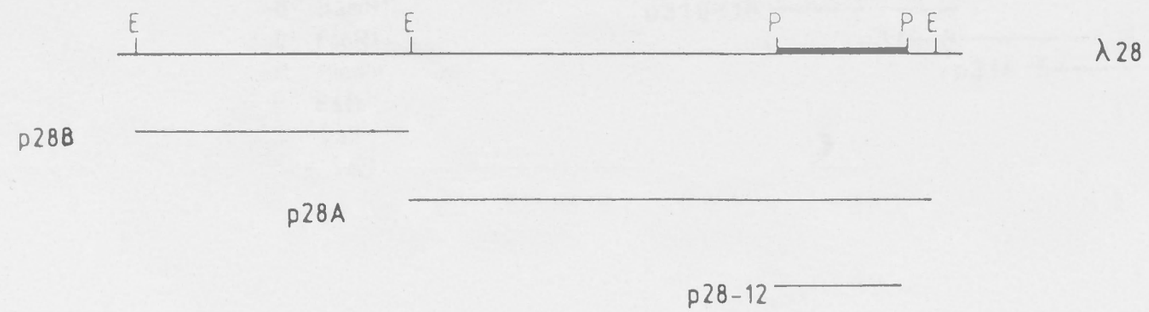
c) G0E6


E E E B B B EB E B
316
p316B6
p316B9
p316C
p48-4
p316D
p316-3A
p316-8AA
E B B B EB E B E H B H HE
48
p48-13
p48-3
p48-11
B BamH1
E EcoR1
H HindIII
P Pst
B E H B H HEPB P H E B
41
p41-20
p41-11
1kb

D) GOE7



E) GOE8



F) GOE9

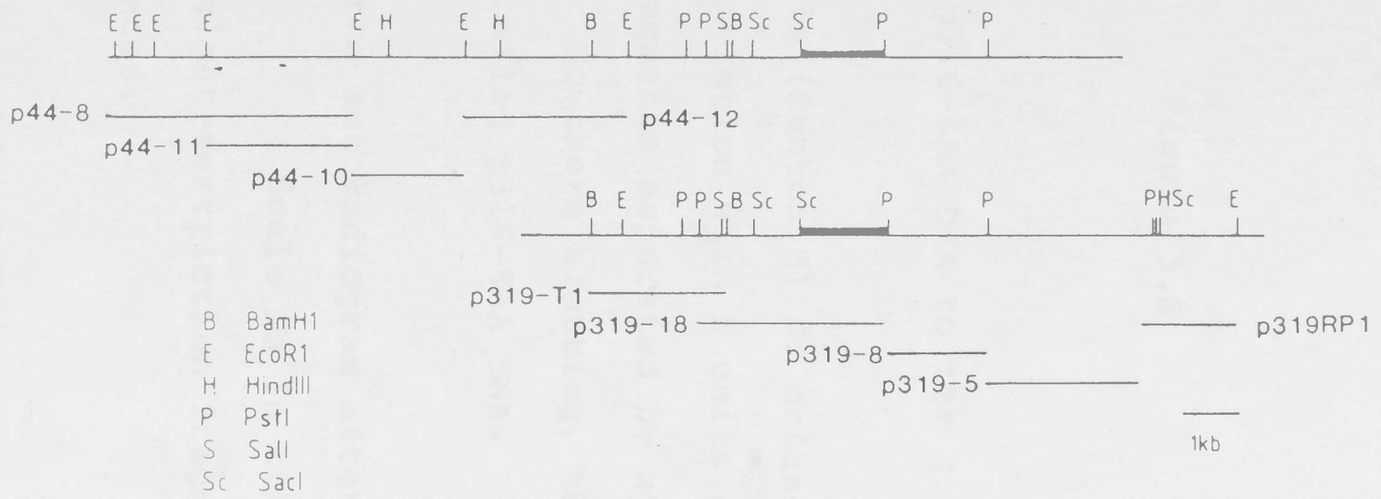

E E E E E H E H B E P P S B Sc Sc P P
p44-8
p44-11
p44-10
p44-12
B E P P S B Sc Sc P P P H Sc E
p319-T1
p319-18
p319-8
p319-5
p319RP1
1kb
B BamH1
E EcoR1
H HindIII
P PstI
S SalI
Sc SacI

Figure 3.8

Hybridisation of p316-8AA DNA to the *D. melanogaster* genome.

5ug of adult female (Canton S) *D. melanogaster* genomic DNA was digested to completion with 5 units of Alu I enzyme and the restriction fragments separated by agarose gel electrophoresis. After Southern blotting, the DNA was probed with radio-actively labelled p316-8AA DNA.

The figure shows the autoradiogram after one week's exposure. Scale is in kilobase pairs (kb). HMW = high molecular weight restriction fragments that hybridise to GOE sequences.

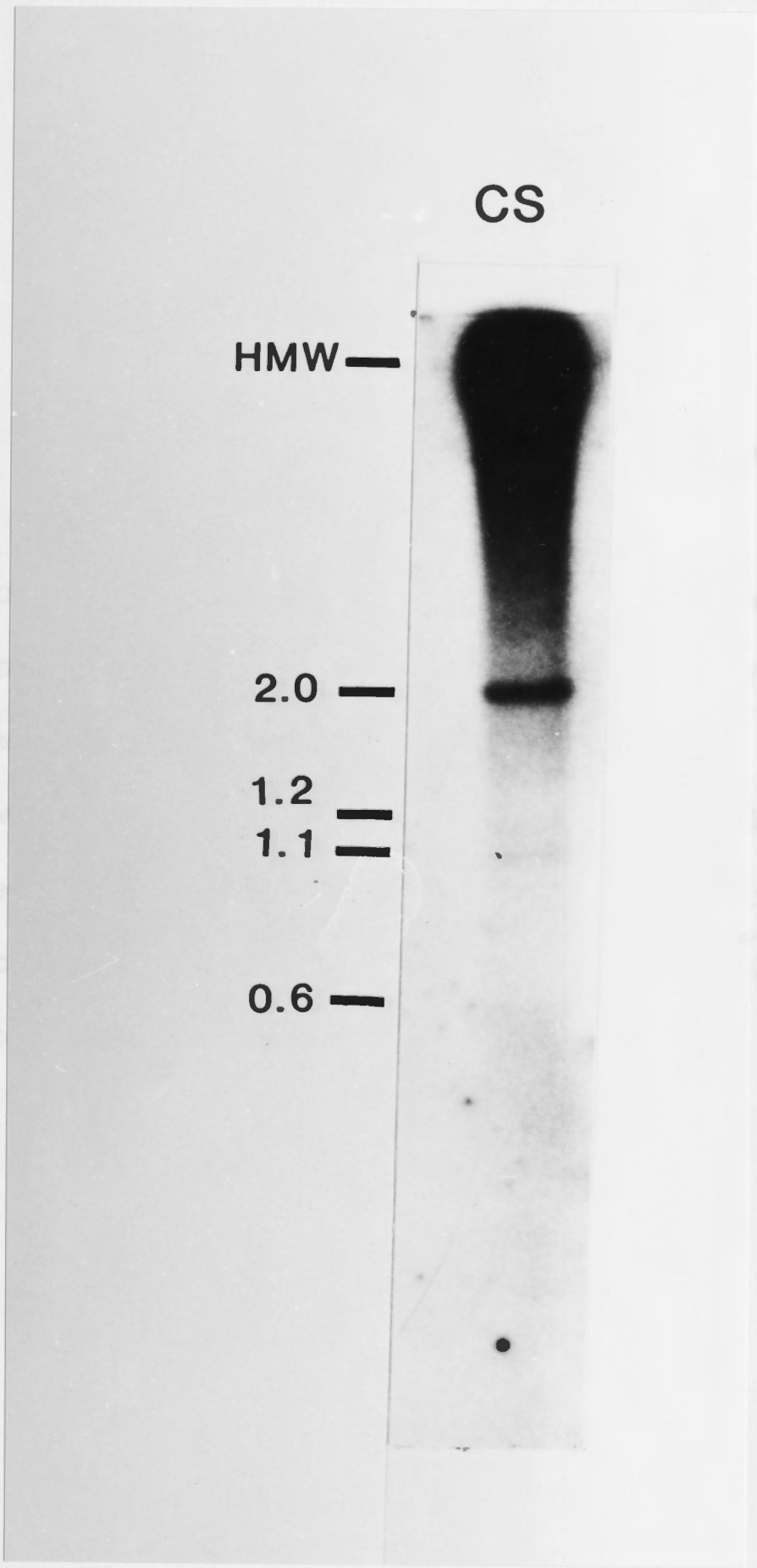

CS
HMW
2.0
1.2
1.1
0.6

Figure 3.9

Hybridisation of single-stranded GOE sequences to the *Drosophila* genome.

Two aliquots of 3ug of *D. melanogaster* adult female DNA were digested to completion with 5 units of Alu I. Fragments were separated on a 1% agarose and Southern blotted. One filter was probed with a radiocatively-labelled single-stranded DNA of the 5'-GATA-3' strand of GOE6. The other was probed with radioactively-labelled single-stranded DNA of the 5'-TATC-3' strand of GOE6.

5'-GATA-3' 5'-TATC-3'

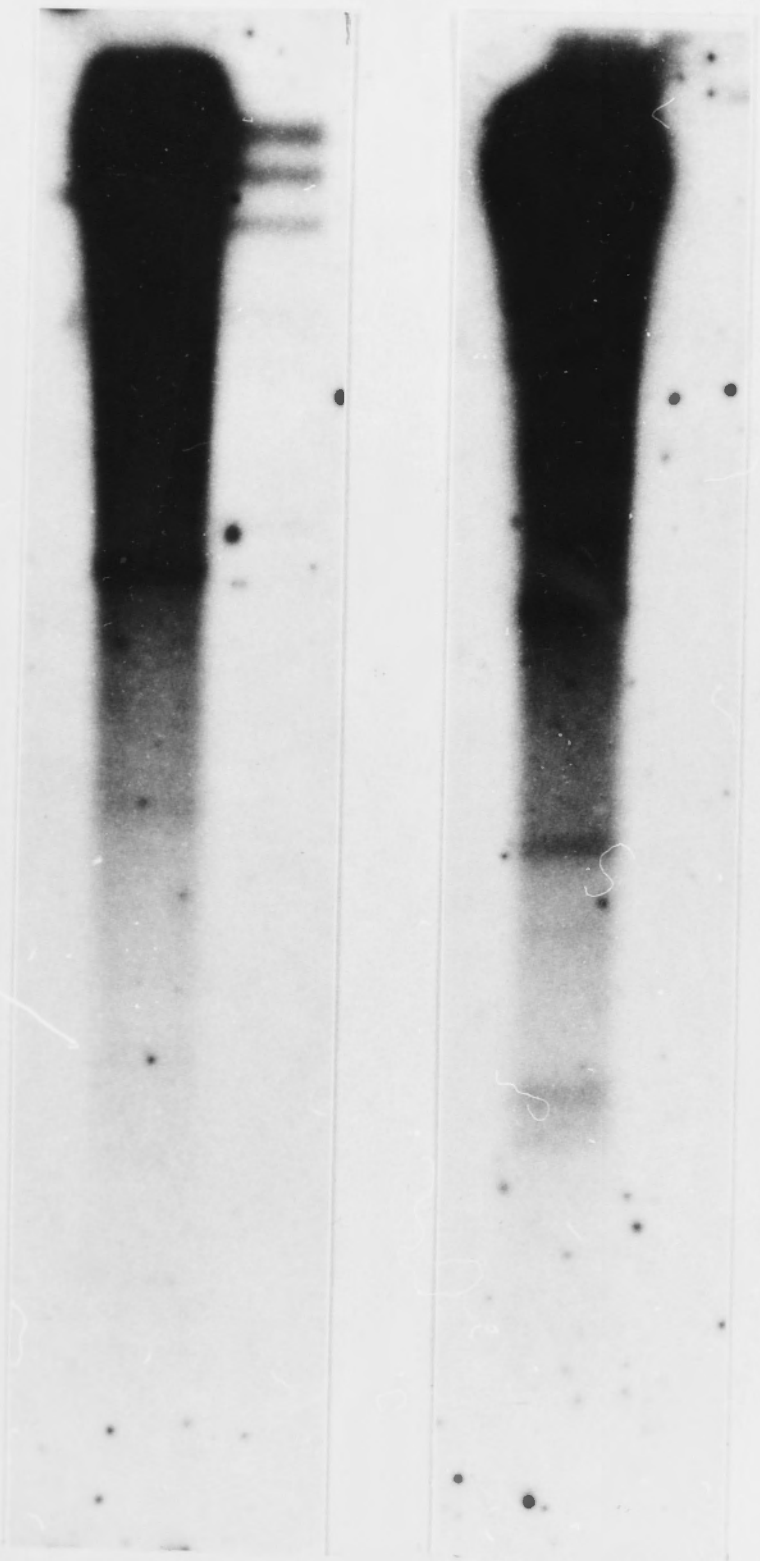


Figure 3.10

Hybridisation of sACl and p47-13 DNAs to male and female *Drosophila* genomes. 2ug of male and female *D. melanogaster* DNAs isolated from adult heads (CS(male) and CS(female), respectively) were digested to completion with 5 units of Eco RI enzyme. Fragments were separated on a 1% agarose gel, Southern blotted and simultaneously probed with radioactively-labelled sACl and p47-13 DNAs.

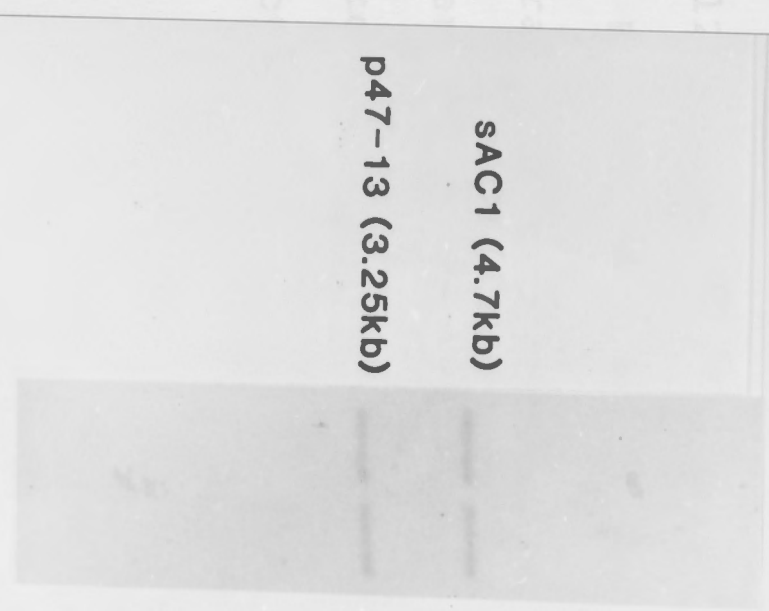
Laser densitometer tracings of the hybridising bands were made of the resulting autoradiogram (shown in the figure) and the relative areas of the corresponding peaks estimated.

MALE (p47-13) : (sACl) = 1 : 1.11

FEMALE (p47-13) : (sACl) = 1 : 1.46

MALE : FEMALE = 1 : 0.76.

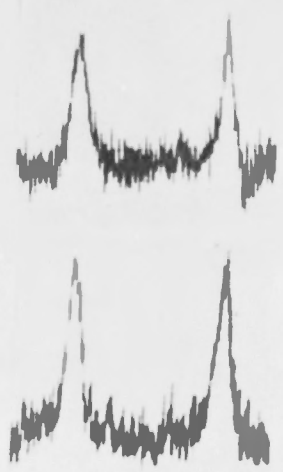
a)



CS(male)

CS(female)

b)



CS(male)

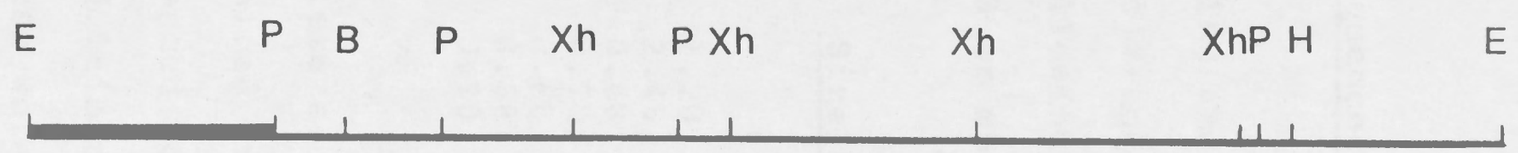
CS(female)

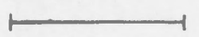
Figure 3.11

Restriction enzyme map of the *Drosophila* insert in the plasmid, pFF12. The plasmid was subcloned from the lambda recombinant, FF4 that had been isolated from a genomic library of the FF strain of *Drosophila melanogaster*. Symbols for the restriction enzyme sites are explained in the key. The region hybridising to p316-8AA (heavy line) lies in a 0.7kb Eco RI-Pst I restriction fragment.

Restriction enzyme map of pFF12

pFF12





500bp

B	Bam H1
E	Eco R1
H	Hind III
P	Pst I
Xh	Xho I

Chapter 4

SEQUENCING OF RESTRICTION ENZYME FRAGMENTS CONTAINING GOE SEQUENCES

4.1 Strategies for cloning GOE sequences into M13 vectors

The M13 system for the subcloning and subsequent sequencing of DNA restriction enzyme fragments has been summarised in section 2.10. The following restriction fragments were subcloned into M13mp8 or mp9 vectors:

GOE sequence	Subclone	Restriction enzyme	Size(kb)	Cloned in:
GOE4	p314-4B	EcoRI-PstI	1.20	both orientations
GOE4	p314-4A	EcoRI-PstI	2.40	one orientation
GOE5	p315-T22	Hae III	0.48	both orientations
GOE5	p315-T22	Sau 3AI	0.70	one orientation
GOE6	p316-8AA	Taq I	0.68	one orientation
GOE6(FF)	pFF3	Taq I	0.68	both orientations
GOE12	pFF12	EcoRI-PstI	0.70	both orientations

Except for GOE4 and GOE6, complete sequence data for these restriction fragments was obtained from the resulting M13 recombinants. p314-4A was sequenced for up to 200 bases from the Pst I site. A number of additional recombinants containing the GOE4 and GOE6 sequences were obtained by subcloning the products of BAL 31 digestions of p314-4B and p316-8AA DNAs into the EcoRI and HincII sites of M13mp8 (see

Materials and Methods section for details). In this way, sequence data for the entire 1.2kb p314-4B and for much of the remainder of the 0.68kb Taq I p316-8AA restriction fragments were obtained.

The strategies by which the various GOE sequences were obtained are summarised in figure 4.1. The extent and direction in which the sequences of the M13 recombinants were read are indicated by the arrows. An example of a sequencing gel is presented in figure 4.2.

The available sequence data for the GOE clones are presented in figure 4.3a-f. Included here is the sequence for the GOE9 sequence that was published by Singh et al. (1984).

4.2 Overview of the GOE sequences

From an initial view of the sequence data, the following general points can be made:

i) All the sequences contain large numbers of the tetranucleotide, GATA*. No other simple sequences are as predominant.

ii) These GATA units are arranged in blocks. The longest block is of 34 units in the GOE6 copy from the FF strain.

* By convention, the repeat unit is referred as GATA, though it can equally well be defined as AGAT, TAGA or ATAG (as it is in Alonso et al. 1983).

iii) The total number of GATA units is not constant between GOE sequences. For example, GOE4 has 64 units whereas GOE5 has only 23.

iv) Adjacent blocks are interrupted usually by sequences of four, or multiples of four, bases. These quadruplets can often be converted to a GATA by a single base change. For example, from position 144 to 159 in GOE4, there is the sequence GATA GATG GATT GATA. The second and third tetranucleotides can be converted to GATAs by replacing G with A at position 151, and T with A at position 155. Occasionally, one needs to postulate that a deletion has occurred in order to maintain this 4 base pair periodicity, e.g. from positions 402 to 412 in GOE4 there is the sequence GATAATAGATA. Insertion of a G residue at position 406 restores the GATA periodicity. Consequently, even interrupted blocks can be converted with few changes into longer, continuous stretches of GATA.

v) Beyond the GATA regions, there appear to be no sequences that are common to all the GOEs.

vi) If there were a transcript of a continuous GATA stretch it would contain no stop codons throughout its length. Theoretically, it could then be translated. This possibility has been expressed elsewhere (Epplen et al., 1983b and Singh et al., 1984).

It can also be seen from the sequences in figure 4.3 that the corresponding regions of the GOE6 and GOE6(FF) sequences are almost identical except for two regions (which are boxed in the figure). These differences are discussed in more detail in Chapter 5. The second GOE sequence derived from the genome of the FF strain (GOE12) is dissimilar to the sequences of the other Canton S GOE sequences. It may be equivalent to either GOE7 or GOE8, which have not yet been sequenced, though as mentioned earlier, the flanking sequences of these three GOE sequences do not cross-hybridise. If GOE12 is the FF equivalent of GOE7 or GOE8, it would have had to have transposed to another location to account for the fact that its flanking sequences do not correspond to those in the Canton S strain.

Figure 4.1

Sequenced regions of restriction enzyme fragments containing GOE sequences. Arrows indicate starting points and direction in which sequences were read from the M13 recombinants. Lengths of the arrows indicate the extents to which sequences were read from single reactions. Heavy lines show the GATA-rich regions in each case.

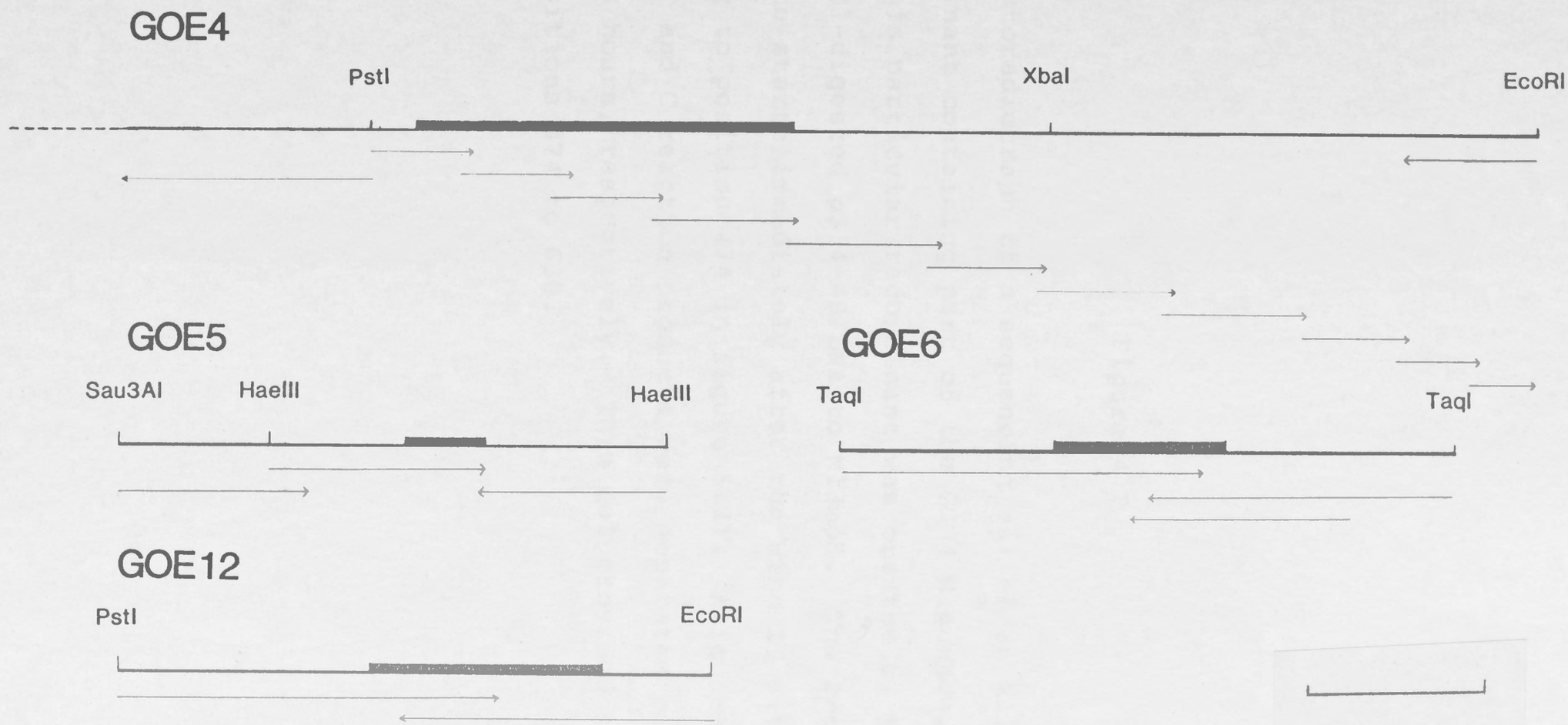

GOE4
PstI
XbaI
EcoRI
GOE5
Sau3AI
HaeIII
HaeIII
GOE6
TaqI
TaqI
GOE12
PstI
EcoRI
200 bp

Figure 4.2

Autoradiograph of a sequencing gel of an M13mp8 recombinant containing part of the GOE4 sequence.

This particular recombinant was created by the ligation of BAL31-digested p314-4B DNA to M13mp8. The *Drosophila* sequence starts immediately after the Hinc II site (corresponding to position 474 in figure 5.2). Aliquots of the four G, A, T and C reaction products were separated on the gel for 3 and 8 hours, respectively. This gel provided sequence data for positions 474 to 630.

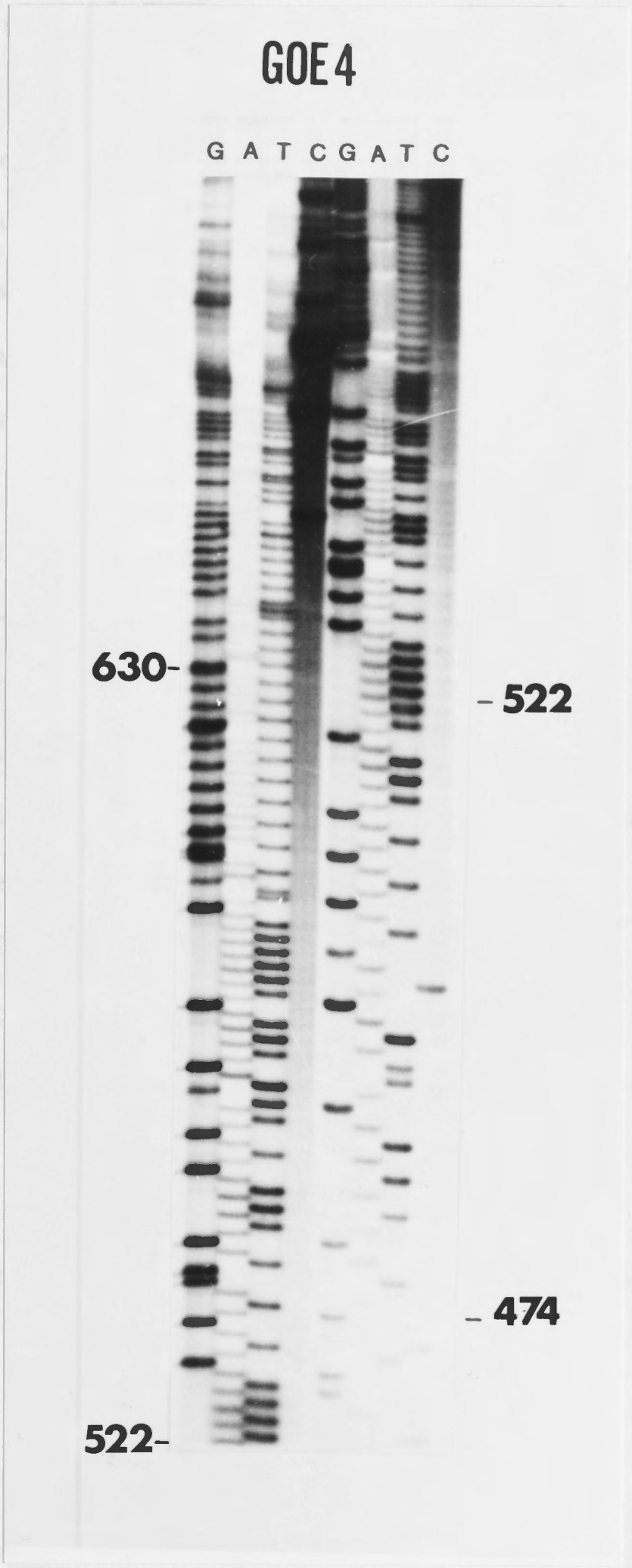

GOE4
G A T C G A T C
630-
522
474
522-

Figure 4.3.

Sequences of restriction enzyme fragments containing GOE elements.

Sequences are shown 5' to 3'. Only one strand containing 5' - GATA - 3' is shown, though in some regions the complementary strand was read from the sequencing gel. Sequences printed in upper case letters correspond to the GATA regions as they are defined in section 5.1.

GOE4 Combination of sequence data from all of p314-4B and part of p314-4A.

GOE5 Combination of sequence data from the 450bp Hae III and 600bp Sau 3AI fragments of p315-T22.

GOE9 Sequence of GOE from the lambda clone 319, published in Singh et al., (1984).

GOE12 Sequence of 0.7kb Eco RI - Pst I fragment from pFF12.

GOE6 Sequence of part of the Taq I fragment from p316-8AA.

GOE6(FF) Sequence of the whole of the 693bp Taq I fragment from pFF3.

Boxes indicate those regions where GOE6 and GOE6(FF) differ.

A,a = adenosine,

C,c = cytosine,

G,g = guanosine,

T,t = thymidine

a) GOE4

```
        10         20         30         40         50         60         70         80         90        100
tctttcggta ttgcgactgc aactagaagc cttgcgtttt cttgccctgg tgtgaattta ttgcttaatc atgcgacttt aaaacggcac ggctggctgg

       110        120        130        140        150        160        170        180        190        200
caggtaaagg cagctgggtt ttggattcgg ttttaacaag gtgagtgctt gtccaactgg acgcgctccg ttttatttt ctggccatat tctacggcat

       210        220        230        240        250        260        270        280        290        300
cttcccggta ggcgtaggtg gtagcaaatg gcctgaaaaa ggacacacaa aaagaaacgg aaggtcattt gtggctgcag ctcagtgtcg tttgttttat

       310        320        330        340        350        360        370        380        390        400
tcgccgagtg gatgatgaat atataatata GATAGATATA TAGATAGATG TATAGATGGT TAGATAGATA GGTAGATAGA TAGATAGATA GATAGATGGT

       410        420        430        440        450        460        470        480        490        500
AAGATAGATA GATAGATAGA TAGATAGATG GTTAGATAGA TAGATAGATA GATAGATAGA TAGATAGATA GATAGATAGA TAGATATATA GATTATAGAC

       510        520        530        540        550        560        570        580        590        600
AGATAGATAG ATAGATATAT AGATATATAT ATATAGATAG ATAGGTAGAT ATATAGATAG ATATGTAGAT ATATAGATAT ATATATATAG ATTGATAGGT

       610        620        630        640        650        660        670        680        690        700
AGATAGATAG ATAGATAGAT GGATAGATAG ATAGGTAATA GATAGATATA TAGATAGATA GATAGATAGA TAGATAGACA GATAATAGAT ATAAAGATAG

       710        720        730        740        750        760        770        780        790        800
ATATAAAGAT AGATAGATAG ATCAACCAAT AGATAGATGT TAGATAGATA GATAgtttga ttgcaacctg ccacgatgtt cgattctggc tggcatgttg

       810        820        830        840        850        860        870        880        890        900
aagttcgatt cgaacaagct ggtgtttgag tccaacaatc ttgggcttaa attcttctta aagtataaaa taacaattcc cttaaaattg agatttagag

       910        920        930        940        950        960        970        980        990       1000
ttttactttt gagttctggc taaacttatt tttattagat aacacatcta cttctcgaac ccttgctctg cctctgcggt cattttcgtg accagattcg

      1010       1020       1030       1040       1050       1060       1070       1080       1090       1100
gaattgtgaa tagttaggtt gtgagctctt ctagacagct gatatgcaaa tatgatagag gttagaattg caccatgttt tgattattta taaatcaaat

      1110       1120       1130       1140       1150       1160       1170       1180       1190       1200
caaagtatta ataatatcat gtgatgatga atttgtcgat ttaaaatgtg aaagtttttg ttaggtatgt aaatagcaaa ctaataacag agggaaattt

      1210       1220       1230       1240       1250       1260       1270       1280       1290       1300
agttccgtgt aaatctttcg acagagagtc ttagacagga tgtcttggaa actaaaagac gtgagaatta agaacaatga ctaaataatt ttaaatgaaa

      1310       1320       1330       1340       1350       1360       1370       1380       1390       1400
tatatccatt aaatatttct tattaatgag atatttaaat aaggatttct ttacacagtt cagtatttta catcgaattt gaatatcact catacgcctg

      1410       1420       1430       1440       1450       1460       1470       1480       1490       1500
tgttaaaaat tatattttca cctggttagg aaggagaaaa actgaacgga ggcaaggcaa catgttcgaa tcgaactgcc gagtctttcg cgcgggtaac

      1510       1520       1530       1540       1550       1560       1570
taactgcgaa cataaagctc agtctaaaaa cgagtttcaa cgcgttcggc tgccgacccg ctcgattttt gaattc
```

b) GOE5

```
        10         20         30         40         50         60         70         80         90        100
gatcaaatat ttggtccact tgcaacgacg tcgtcaaagt ttgtttgacc ggcgtaaata aacaatgtgc cttagtcatg ggtggcgaca ataaaatcac

       110        120        130        140        150        160        170        180        190        200
gacacacaag aacagccaca gaaacagcaa gctgctaaag ctgtaagtgc aacaggcaag aaaataaaca ggcctaaaaa tattggataa aaaaccagaa

       210        220        230        240        250        260        270        280        290        300
ggaaaatgat aagtattcaa tccattgaat ccacttgtaa tactaccact gctttagctc ttatcatgca ttgtattaag gcaatatacc ccagttactt

       310        320        330        340        350        360        370        380        390        400
tagacggcat acatatatGA TAGACAGATA GATAGATACA TAGATAGATA GATAGATAGA TAGATAGATA GATAGATAGA TAGATAGATA GATAGATAGA

       410        420        430        440        450        460        470        480        490        500
TAGATAGATA tatagatttg actatctttt agatagacca tctttagacc atatattaca ttagcaattg atctttggtc agtatggcta tctcttaact

       510        520        530        540        550        560        570        580        590        600
agttatgacc tactacattt tgtgcgcatg cattggcgac ataagctcga tacgaacagc atgtgacata gacaacttgc agcgcgagat gcagttgatc

       610
cgccccgaac ttggcc
```

c) GOE9

```
        10         20         30         40         50         60         70         80         90        100
tatatatata GATAGATAGA TAGATAGAGA GATAGATAGA TAGATAGATA GATAGATAGA TAGATAGATA GATAGATAGA TAGATAGATA GAGAGAGAGA

       110        120        130        140        150        160        170        180        190        200
TAGATAGATA GATACATAGA TAGATAGATA GATAGATAGA TAGATAGATA GATAGATAGA TGGATAAATA GATAGATAGA TAgattccta tgaatgatcc

       210        220        230        240        250        260        270        280        290        300
tcctaaactt atacttacac ccatatcata cactatatga tgtttttatc aatcaattgt catccttata acgttatatg ccttcccatt tggctacatt

       310        320        330        340        350        360        370        380        390        400
aacaaaaatg ttctgtacat ttaataatac tatgcaattt tcaactaatg ctaataaatt ttcctcggtg tagataaacg gattggagtg ggcgacgctt

       410        420        430        440        450        460        470        480        490
atggctcaaa tggaaacatt tcgcgagctt gtcaaacaat cgaatttaat ttatgatttc ctgcgccaac agtatagaca ctgcatgtgg tagacccc
```

d) GOE12

```
        10         20         30         40         50         60         70         80         90        100
ctgcaggtac ccaaacacct ttatttattt tgtgagcttt aatttatatt tatgcaaata agatggccgt gactgctcaa caaatataca tgattatttg

       110        120        130        140        150        160        170        180        190        200
taaatataaa cttacttggc tgccttaata aggtagcaaa actgataaca gtgtgacaga acacccattc ctgatttcta attattccgt tacacttaga

       210        220        230        240        250        260        270        280        290        300
gaaaaaatgt atactttatt cgttgaaagt atgcatcttg aagatgaagg ctattcacgt ctctgttgat ttgtcaaaat aGATAGATAG ATAGATAGAT

       310        320        330        340        350        360        370        380        390        400
AGATAGATAT ATAGATAGAT AGATAGATAG ATAGATATAT AGATAGATAG ATAGATAGAT AGATGGATAG ATAGATAGAT AGATAGATAG ATAGATAGTT

       410        420        430        440        450        460        470        480        490        500
AGATAGATAG ATAGATAGAT AGATAGATAG ATAGATAGAT AGATAGATAT ATATATAGAT AGATAGATAG ATAGATAGAT AGATAGATAG ATAGATAGAT

       510        520        530        540        550        560        570        580        590        600
AGATAGATAG ATTGATAGAT AGATAGATAG ATAGATAGAT Agatggtcga ctacttaaat attaagtaaa ttaaatatta ttataattta agtattatta

       610        620        630        640        650        660
tattaattat cctgaagctg tacaaatgga atctctagga attgcataca atgtcagaat tc
```

e) GOE6

```
        10         20         30         40         50         60         70         80         90        100
tcgagcgacg aaagactatt gatttcagaa aacatccgaa ccggtatctc tggcactttg tatcaatgaa ctgaaccaaa gatcaagata catttgtgcg

       110        120        130        140        150        160        170        180        190        200
tttgaatgtc tgctccgtct gtgtattttg ttttttttt tttactcaac agcaaattgt ttaaataaat aaaaagacaa gtggatggtg cgctcattga

       210        220        230        240        250        260        270        280        290        300
tatttcaccc aaaaacgatt ttgagaaagc ataaatagaa GATAGATAGA TTGATAGATA GATTGATAGA TAGACTGATA GATAGATAGA TAGATAGATA

       310        320        330        340        350        360        370        380        390        400
GATAGATAGA TAGATAGATA GATAGATAGA TAGATAGATA GATAGATAGA TAGATAGATA GATAGATAGA TAGATAGATA GATAGATAGA TAGATAGATA

       410        420        430        440        450        460        470        480        490        500
GATAGATAGA TAGAATAGAT AGAGATAaat tgcacatgct tcaataattt ctatctaaga tcggcgactc acctgaaagc tctcctgcat ccggccaccc

       510        520        530        540        550        560        570
gactgcgtac ttgacgcacc gcccgatgtt gacactatgg tcactggtgc ttccgatccg gacgcagcgc tc
```

f) GOE6(FF)

```
        10         20         30         40         50         60         70         80         90        100
tcgagcgacg aaagactatt gatttcagaa aacatccgaa ccggtatctc tggcactttg tatcaatgaa ctgaaccaaa gatcaagata catttgtgcg

       110        120        130        140        150        160        170        180        190        200
tttgaatgtc tgctccgtct gtgtatttgg ttttttttt tactcaacag caaattgttt aaataaataa aaagacaagt ggatggtgcg ctcattgata

       210        220        230        240        250        260        270        280        290        300
tttcacccaa aaacgatttt gagaaagcat aaatagaaGA TAGATAGATT GATAGATAGA TTGATAGATA GACTGATAGA TAGATAGATA GATAGATAGA

       310        320        330        340        350        360        370        380        390        400
TAGATAGATA GATAGATAGA TAGATAGATA GATAGATAGA TAGATAGATA GATAGATAGA TAGATAGATA GATAGATAGA TAGATAGATA GATAGATAGA

       410        420        430        440        450        460        470        480        490        500
TAGATAGATA GATAGAATAG ATAGAGATAa attgcacatg cttcaataat ttctatctaa gatcggcgac tcacctgaaa gctctcctgc atccggccac

       510        520        530        540        550        560        570        580        590        600
ccgactgcgt acttgacgca ccgcccgatg ttgacactat ggtcactggt gcttccgatc cggacgcagc gctccagatt gttgcgcatc ttcgtatgac

       610        620        630        640        650        660        670        680        690
gctgctgctg ctgagcagaa tgcccgctcc aaggagcaat agttgcgatc actcgctccg gtgctcctcc tcgttcgccg gcggttcakt cga
```

Chapter 5

ANALYSIS OF THE GOE SEQUENCES

5.1 Definition of the GOE sequence

GOE sequences were previously rather loosely defined as GATA-rich sequences that are present in the sex-chromosome associated snake satellite DNAs and in a number of eukaryote genomes. A more accurate definition for this family of repeated sequences can now be obtained by comparing the *Drosophila* GOE sequences that were presented in Chapter 4. The GATA-rich regions do not consist purely of contiguous GATA tetranucleotides, nor are they of the same length and it is not immediately obvious whether the non-GATA units are integral components of the GATA region. It is possible that there is no definite start and finish to the GOE sequence so that the it would be best described as a local concentration of GATA units. In order to visualise the arrangement of GATA sequences (and of their single base variants), they are presented schematically in figure 5.1. The method used to visualise the GOE sequences assumes that the sequenced regions were initially derived from a continuous stretch of GATA units, and so they were defined in terms of 4 base pair blocks. GATA units were first identified (fully shaded), followed by single base variants of the canonical GATA unit

(half shaded). Section 5.5 describes these single base variants more fully. The single base deletions, GTA and GAA, were also identified and placed in this latter class (though *G*GTA and GAA*A* would already have been included). The other two possible deletion variants, ATA and GAT, are automatically included in variants of the form, *N*ATA and GAT*N*, respectively, where *N* represents any of the four nucleotides, A,C,G or T.

The remainder of each sequence was then divided into blocks of 4 nucleotides (unshaded) wherever possible. This left several islands of 3, 2 or 1 nucleotides, surrounded by previously defined blocks. These needed to be incorporated into the structure, in order to preserve the integrity and length of the sequence. Each 3 nucleotide block was classed as a four nucleotide block and, to compensate for this, a corresponding number of single nucleotides were deleted. Every second dinucleotide was then assumed to be a four nucleotide block and the remainder were deleted from the sequence. In this way, the original lengths of the sequences were maintained. However, a number of 'deletions' and 'insertions' have had to be assumed in order to maintain an artificially-imposed four nucleotide periodicity throughout. Consequently, more apparent single base variants have been assigned than if each sequence had been divided into four nucleotide blocks *before* identifying GATAs and their variants. Single base changes, deletions and insertions within the GATA regions are, however, now more clearly defined (see section 5.5, for examples).

Some of the apparent single base variants outside the GATA region will arise simply as random assortments of a nucleotide sequence (any 4 base sequence will occur every 256 nucleotides, on average). To obtain a visual measure of this effect, the same method as used above was applied to a 1000bp randomly generated sequence containing equal proportions of the four nucleotides, 300 bases of which are shown in figure 5.1. Included in figure 5.1 are similar representations of other, published GOE-like sequences and their surroundings.

For the *Drosophila* GOE elements at least, the GATA units do seem to be localised. Except for GOE5 and GOE6, the 5' and 3' ends of the GATA regions are defined, respectively, as the first and last GATA* doublets that are encountered on traversing each sequence. In the case of GOE5, the first doublet is preceded by the sequence, GATAGACA. A single base change can convert this to a GATA doublet so this sequence was included in the GATA region of GOE5. For GOE6, the last GATA doublet is followed by the sequence, GAATAGATAGAGATA. Two deletions (at the underlined positions) will convert this to a GATA triplet and this sequence also was included in the GATA region of GOE6. This definition incorporates almost all GATA units into the GATA region, except for a few single GATAs that lie no closer than 20 bases from the main body. More than 60% of the GATA regions is composed of GATA units and the

* The probability of finding a single GATA unit in a randomised sequence composed of equal proportions of A,C,G and T residues is 1 in 256. The probability of finding a *doublet* is 1 in 65,500.

remainder consists predominately of single base variants of the GATA unit itself. The limits of the GATA regions, as defined here, are numbered in figure 5.1 and the regions themselves are printed in upper case letters in figures 4.3a-f.

5.2 Search for sequence similarity beyond the GATA regions.

5.2.1 Comparison of the sequences immediately adjacent to the GATA regions.

Immediately 5' of the GATA regions, there are short (10 nucleotide) AT-rich sequences. These occur as tandemly arrayed TA dinucleotides in the case of GOE4, 5 and 9. At the 3' ends, most of the GATA regions are followed by a single base variant of GATA, though these are not identical to each other. Nucleotides shared by three or more GOE elements are boxed in figure 5.2 to illustrate the extent of homology. Up to 10 bases at the 5' end and 5 bases at the 3' end of the GATA regions could be regarded as similar (that is, matches occur in at least three of the five sequences), though at the 5' end it was necessary to introduce insertions into the GOE5 and GOE6 elements and a deletion in the GOE4 element to obtain a maximum number of matches. Even if these short regions of similarity do mark the real limits of the GOE sequences, they are not clearly defined and are not considered as part of the GATA region for the analysis that follows.

To search for any similarity between the flanking regions on a broader scale, two computer-assisted methods were employed*, as described below.

5.2.2 Comparison of the sequences beyond the GATA regions by the Dayhoff alignment method.

The Dayhoff program uses a method developed by Needleman and Wunsch (1970) that was originally intended to align amino acid sequences. Two sequences are aligned so that the greatest number of matches between the nucleotides is obtained. Gaps in either sequence can be inserted, though a penalty may be accrued for each gap if desired (that is, one can select against the occurrence of deletions/insertions that may otherwise be required to provide a maximum alignment). For the GATA regions and their flanking sequences, no such penalty was applied, effectively allowing for a greater degree of diversity between the sequences.

To guage the significance of the best possible alignment, the highest number of matches (R) is compared to the mean number of matches (M) that is obtained when 20 different randomly generated sequences are aligned (each composed of same proportion of nucleotides as the original sequences). The *alignment score* is the difference, (R-M) divided by the standard deviation in the number of matches between the

* All computer-assisted analyses were carried out on a VAX 11750 ng under VMS) system, using the SEQUENCE package. Except for the Dayhoff alignment program, all procedures of the SEQUENCE package that were used here were created at RSBS.

randomised sequences - that is, it is the number of standard deviations of the mean by which the observed alignment deviates from the expected mean. Alignment scores of less than 2.0 show there to be a greater than 5% probability that the observed alignment arose by chance. Scores of more than 2.0 were therefore taken here to indicate that the observed alignment is significant at the 5% level.

Alignment tests were carried out for the 200bp lying immediately 5' and 3' to the GATA regions.

The results are summarised in table 5.1. As expected, the GATA regions themselves show good similarity whereas most of the flanking regions show very little. Significant similarity is seen between the 3' flanking regions of GOE5 and GOE12 only, although the scores for the alignments of the 3' flanking regions of GOE5 with GOE6 and of GOE5 with GOE9 are greater than 1.5. However, alignments between the 3' flanking regions of GOE6, GOE9 and GOE12 are not significant. This result could be interpreted as showing that the 3' flanking regions of GOEs 6, 9 and 12 each have different components which are however all present in the corresponding region of GOE5. A second method of comparison was therefore employed to see if this also will reveal the apparent similarities between the various 3' flanking regions.

* Strictly speaking, scores of less than -2.0 are also significant. However, these cases suggest that there is less similarity between sequences than would be expected by chance, and are not germane to a search for significant positive similarities

Table 5.1. Alignment scores for the flanking sequences of the *Drosophila* GATA regions

5' flankers

	GOE4	GOE5	GOE6	GOE12
GOE4	-	-1.89ns	-0.86ns	-1.52ns
GOE5	-1.89ns	-	-2.75*	0.62ns
GOE6	-0.86ns	-2.75*	-	-3.23*
GOE12	-1.52ns	0.62ns	-3.23*	-

3' flankers

	GOE4	GOE5	GOE6	GOE9	GOE12
GOE4	-	-1.57ns	-1.08ns	0.60ns	-1.54ns
GOE5	-1.57ns	-	1.76ns	1.66ns	2.34*
GOE6	-1.08ns	1.76ns	-	-0.44ns	1.32ns
GOE9	0.60ns	1.66ns	-0.44ns	-	0.97ns
GOE12	-1.54ns	2.34*	1.32ns	0.97ns	-

ns = not significant.

* = <5% probability of the alignment occurring by chance.

5.2.3 Comparison of the sequences beyond the GATA regions by the dot matrix diagram method.

In the dot matrix diagram option, the two sequences to be compared form a two-dimensional matrix. The nucleotides at each position of one sequence are compared to the nucleotides at every position in the other sequence. If a defined number (4, in this case) of nucleotides in a row are matched, their positions in the matrix are marked. Two identical sequences will produce a continuous, diagonal line through the matrix.

Each GOE sequence was compared to the others in this way and examples of the results are shown in figure 5.3. It can be seen that no extensive similarity is revealed between the flanking regions of GOE6 and those of GOE5 and GOE12. Owing to the tandem repetition of GATAs in all the GOE elements, matches are found at all positions within each GATA region. Therefore the GATA regions appear as fully shaded blocks in the figure. Beyond the GATA regions, no matches of 10 or more bases were revealed between any pair of GOE elements, showing that there is little sequence similarity between the 5' or 3' flanking regions.

These flanking regions may have a structure distinct from randomly assorted sequences, and so the number of matches between all pairs of flanking regions were compared to the number or matches between two randomly generated sequences. For 200 nucleotides of randomly generated sequences containing A, C, G and T nucleotides in the same proportions as are found in the combined flanking regions (A = 0.31, C = 0.19, G =

0.19, T = 0.31), it was found that the mean number of 4bp matches is 136.5 ± 14.5. The significance of the observed number of matches for each pairwise comparison was tested against this mean, and the results are summarised in table 5.2.

There are several pairwise comparisons which show a significantly high number of matches, both between 5' and between 3' flanking regions*. However, none of the these pairwise comparisons correspond to those that showed significant similarity in the Dayhoff alignment test.

The mean number of matches between two randomly generated sequences increases as the (A + T) content deviates from 0.5 (Moore et al., 1984). This could explain, for example, the very high number of 4bp matches between the 3' flanking regions of GOE9 and GOE12, which have (A + T) contents of 0.65 and 0.75, respectively.

5.2.4 Summary.

Both the alignment and matrix procedures show there to be no consistent nucleotide sequence similarity between the 5' and 3' sequences flanking the GATA regions. The only similarity lies within the GATA regions, with the possible inclusion of the 5-10 nucleotides lying immediately adjacent. The term *GOE element* will be used to refer to these GATA regions as they were defined in section 5.1. The GOE elements have been printed in upper case letters in figure 4.3a-f.

* A significant mismatch was found between the 3' flanking regions of GOE4 and GOE6, but was not further considered (see argument in the footnote on page 98).

Table 5.2 Number of 4bp matches between the flanking sequences of the *Drosophila* GATA regions.

5' FLANKING REGION

	GOE4	GOE5	GOE6	GOE12
GOE4	-	123ns	155ns	144ns
GOE5	123ns	-	163*	167*
GOE6	155ns	163*	-	134ns
GOE12	144ns	167*	134ns	-

3' FLANKING REGION

	GOE4	GOE5	GOE6	GOE9	GOE12
GOE4	-	145ns	106*	161*	170*
GOE5	145ns	-	120ns	150ns	142ns
GOE6	106*	120ns	-	120ns	112ns
GOE9	161*	150ns	120ns	-	202*
GOE12	170*	142ns	112ns	202*	-

Expected no. of matches = 136.5 ±14.5.

ns = not significant. * = significant.

5.3 (A + T) content of the flanking sequences

Although the regions flanking the GOE elements show no sequence homology with each other, their overall nucleotide composition may be distinct from the bulk of the genomic DNA. The mean (A + T) contents of the flanking sequences are; for the 5' flanking region 62.25% ± 5.6 and for the 3' flanking region 63.9% ± 7.8. These values are similar to the average (A + T) content for the *Drosophila* genome as a whole (= 58-60%, CRC Handbook of Biochemistry, 1968).

5.4 Identification of single base variants of the canonical GATA sequence within and around the GOE elements

Each of the GOE elements described contains a proportion of single base variants of the canonical GATA tetra-nucleotide. These variants may have arisen by chance as a result of single base mutations within what was originally a pure, tandem array of GATAs, or they may have a structural role in the function of the GOE element and have thus been maintained. The former possibility would be more likely if it could be shown that the occurence and type of these variants are no different from that expected from a randomly generated sequence. To resolve these two possibilities, the number and types of variants were estimated for the GOE elements, for their surroundings and for randomly generated sequences.

The possible single base variants of GATA are:

Nucleotide (N) substitution	NATA	GNTA	GANA	GATN
A	AATA	-	GAAA	-
C	CATA	GCTA	GACA	GATC
G	-	GGTA	GAGA	GATG
T	TATA	GTTA	-	GATT
Nucleotide deletion	ATA	G TA	GA A	GAT

There are a number of single base insertions that should also be considered:

Nucleotide (N) inserted	GNATA	GANTA	GATNA
A	GAATA	GAATA	GATAA
C	GCATA	GACTA	GATCA
G	GGATA	GAGTA	GATGA
T	GTATA	GATTA	GATTA

Single base deletions and insertions can be unequivocally identified within the GOE elements because one has a natural internal four base periodicity in which to place them. (Only one example of an insertion has been found. At the end of GOE6, lies the sequence 140 GATAGAATAGATA 152, which could have been produced by the insertion of an A residue within the middle GATA unit).

Single base deletions and insertions in the flanking regions cannot be identified so unequivocally. The deletion

variants, ATA and GAT, will automatically be incorporated into the class of variants, NATA and GATN as they must lie next to one of the four bases. Of the insertions, only GACTA and GAGTA are distinguishable, for the remainder are of the form, GNATA and GATNA. However, insertions of this form could be assumed if it was shown, for example, that the number of GCATA sequences is greater than the product of the number of CATA sequences and the fraction of G nucleotides present in the sequence as a whole. The excess could be regarded as genuine insertions for the purposes of this analysis, though one could not say which of the 5 base sequences should be the insertion variant.

The observed and expected values for the numbers of the various insertion sequences in the flanking regions of the five GOE elements are presented in table 5.3. The values for the fraction of A and G bases reflect their relative proportions in the flanking sequences as a whole. Only in one instance does the observed number of 'insertions' exceed the expected number. There are four GATGAs in GOE4 where two were expected, but these are incorporated into two sequences of the form, GATGATGA and so are not equivalent to genuine insertions. In conclusion, except for the GACTA and GAGTA sequences, the possibility of insertion variants can be ignored.

Table 5.3. Comparison of observed no. of insertions (O) to expected no. of insertions (E)

	GOE4 (27%A, 22% G)			GOE5 (32% A, 18% G)			GOE6 (26% A, 23% G)			GOE9 (31% A, 14% G)			GOE12 (35% A, 14% G)		
Insertion seq.	(O)	(E)	O-E>1?	(O)	(E)	O-E>1?	(O)	(E)	O-E>1?	(O)	(E)	O-E>1?	(O)	(E)	O-E>1?
GAATA	3	3.08	No	0	1.26	No	0	1.15	No	0	0.42	No	0	0.98	No
GCATA	0	0.66	No	1	0.90	No	1	0.23	No	0	0.28	No	1	0.14	No
GGATA	0	0.88	No	1	0.90	No	0	0.46	No	0	0.14	No	0	0.14	No
GTATA	1	1.54	No	0	0.72	No	0	0.00	No	1	1.26	No	1	0.84	No
GACTA	1	1.12	No	1	0.60	No	1	0.51	No	0	0.31	No	1	0.38	No
GAGTA	0	1.47	No	0	0.54	No	0	0.44	No	0	0.22	No	0	0.39	No
GATTA	1	2.70	No	0	0.32	No	0	0.78	No	0	0.93	No	1	1.05	No
GATAA	1	1.08	No	2	1.60	No	0	0.46	No	1	0.31	No	1	0.33	No
GATCA	0	0.00	No	1	0.96	No	1	1.30	No	0	0.31	No	0	0.00	No
GATGA	4	1.62	Yes	0	0.32	No	0	0.52	No	0	0.31	No	1	1.05	No

The numbers and types of GATA variants are displayed in table 5.4. In GOE4 and GOE5, there are a number of variants that require two or more changes to convert them to a GATA. Each change is included in the tables. The 5', GOE and 3' regions have been treated separately. Values for all the GOE elements combined are also listed.

5.5 Analysis of GATA variants in the flanking regions

Though the flanking regions do not show significant sequence similarity, it has not been shown whether the numbers, distribution and types of variants are significantly different from what can be expected from a randomly generated sequence. To test this the numbers of GATA variants in units of 20 bases along the sequenced regions were counted and the results presented in the form of histograms in figure 5.4. There does not appear to be any localised distribution of the variants when compared to a random sequence. The observed distribution was compared to that expected as follows:

There are 13 four base GATA and variant sequences and 2 three base variants ('deletions') that were originally searched for. The probability of finding any four base sequence is 1 in 256 (assuming equal proportions of four nucleotides), and 1 in 64 for each three base sequence (some of the three base variants will be included in the GGTA and GAAA classes). The probability of finding any of the

Table 5.4 GATA variants in the GOE-containing sequences

SEQUENCE OF GOE4

Region	Replacement	Nucleotide replaced				
		G	A	T	A	
5' flanking region n=325	A	1	0	2	0	3
	C	1	0	1	0	2
	G	0	4	0	2	6
	T	3	0	0	1	4
	total replacements	5	4	3	3	15
	deletions	0	1	3	0	4
	total changes	5	5	6	3	19

	Replacement	Nucleotide replaced				
		G	A	T	A	
GOE element n=425	A	2	0	3	0	5
	C	0	0	3	2	5
	G	0	5	0	5	10
	T	14	4	0	2	20
	total replacements	16	9	6	9	40
	deletions	5	2	0	1	8
	total changes	21	11	6	10	48

	Replacement	Nucleotide replaced				
		G	A	T	A	
3' flanking region n=822	A	11	0	5	0	16
	C	2	1	3	0	6
	G	0	2	3	3	8
	T	4	5	0	9	18
	total replacements	17	8	11	12	48
	deletions	0	4	13	0	18
	total changes	13	12	23	12	60

Table 5.4 (contd.)

SEQUENCE OF GOE5

Region	Replacement	Nucleotide replaced				
		G	A	T	A	
5' flanking region n=318	A	7	0	2	0	9
	C	2	1	3	0	6
	G	0	0	0	1	1
	T	1	1	0	0	2
	total replacements	10	2	5	1	18
	deletions	1	5	3	0	9
	total changes	11	7	8	1	27

	Replacement	Nucleotide replaced				
		G	A	T	A	
GOE element n=92	A	0	0	0	0	0
	C	1	0	1	0	2
	G	0	0	0	0	0
	T	0	0	0	0	0
	total replacements	1	0	1	0	2
	deletions	0	0	0	0	0
	total changes	1	0	1	0	2

	Replacement	Nucleotide replaced				
		G	A	T	A	
3' flanking region n=206	A	0	0	0	0	0
	C	3	1	3	2	9
	G	0	0	1	1	2
	T	3	1	0	1	5
	total replacements	6	2	4	4	16
	deletions	0	1	1	0	2
	total changes	6	3	5	4	18

Table 5.4 (contd.)

SEQUENCE OF GOE6

Region	Replacement	Nucleotide replaced				
		G	A	T	A	
	A	3	0	3	0	6
5' flanking	C	1	0	1	1	3
region	G	0	1	1	1	3
n=240	T	0	0	0	2	2
	total replacements	4	1	5	4	14
	deletions	0	2	4	0	6
	total changes	4	3	9	4	20

	Replacement	Nucleotide replaced				
		G	A	T	A	
	A	0	0	0	0	0
GOE element	C	0	0	1	0	1
n=187	G	0	0	0	0	0
	T	0	0	0	3	3
	total replacements	0	0	1	3	4
	deletions	0	0	1	1	2
	total changes	0	0	2	4	6

	Replacement	Nucleotide replaced				
		G	A	T	A	
	A	2	0	1	0	3
3' flanking	C	0	0	1	4	5
region	G	0	0	0	1	1
n=145	T	0	0	0	1	1
	total replacements	2	0	2	6	10
	deletions	0	1	0	0	1
	total changes	2	1	2	6	11

Table 5.4 (contd.)

SEQUENCE OF GOE9

Region	Replacement	Nucleotide replaced				
		G	A	T	A	
GOE element n=180	A	1	0	0	0	1
	C	1	0	0	0	1
	G	0	0	3	1	4
	T	0	0	0	0	0
	total replacements	2	0	3	1	6
	deletions	0	0	0	0	0
	total changes	2	0	3	1	6

Region	Replacement	Nucleotide replaced				
		G	A	T	A	
3' flanking region n=316	A	2	0	1	0	3
	C	2	2	1	1	6
	G	0	1	0	1	2
	T	5	1	0	3	9
	total replacements	9	4	2	5	20
	deletions	1	2	2	0	5
	total changes	10	6	4	5	25

Table 5.4 (contd.)

SEQUENCE OF GOE12

Region	Replacement	Nucleotide replaced				
		G	A	T	A	
	A	5	0	2	0	7
5' flanking	C	0	1	1	0	2
region	G	0	2	1	2	5
n=218	T	4	1	0	3	8
	total replacements	9	4	4	5	22
	deletions	0	2	2	0	4
	total changes	9	6	6	5	26

	Replacement	Nucleotide replaced				
		G	A	T	A	
	A	0	0	0	0	0
GOE element	C	0	0	0	0	0
n=260	G	0	0	0	1	1
	T	4	1	0	1	6
	total replacements	4	1	0	2	7
	deletions	0	0	0	0	0
	total changes	4	1	0	2	7

	Replacement	Nucleotide replaced				
		G	A	T	A	
	A	2	0	0	0	2
3' flanking	C	1	0	0	0	1
region	G	0	0	0	1	1
n=122	T	2	0	0	0	2
	total replacements	5	0	0	1	6
	deletions	0	3	4	0	7
	total changes	5	3	4	1	13

Table 5.4 (contd.)

TOTALS FOR ALL GOE AND FLANKING SEQUENCES

Region	Replacement	Nucleotide replaced				
		G	A	T	A	
	A	16	0	9	0	25
5' flanking	C	4	2	6	1	13
regions	G	0	7	2	6	15
n=1160	T	8	2	0	6	16
	total replacements	28	13	17	13	71
	deletions	0	10	12	0	22
	total changes	28	23	29	13	93

	Replacement	Nucleotide replaced				
		G	A	T	A	
	A	3	0	3	0	6
GOE elements	C	1	0	5	0	6
n=1143	G	0	5	3	7	15
	T	18	5	0	6	29
	total replacements	22	10	11	13	56
	deletions	5	2	1	2	10
	total changes	27	12	12	15	66

Region	Replacement	Nucleotide replaced				
		G	A	T	A	
	A	17	0	7	0	24
3' flanking	C	8	4	8	7	27
regions	G	0	3	4	7	14
n=1729	T	18	7	0	14	39
	total replacements	43	14	19	28	104
	deletions	0	12	21	0	33
	total changes	43	26	40	28	137

'variants' is: 13/256 + (2/64 - 2/256) = 0.074. Put another way, there should be 1.48 variants in every 20 base segment.

The observed distributions of variants in the flanking regions give a good fit to this expected distribution based on a random assortment of bases,

i.e.

GOE4	$P(chi^2 = 16, 12 \text{ d.f.}) > 0.1$	ns
GOE5	$P(chi^2 = 4, 5 \text{ d.f.}) > 0.8$	ns
GOE6	$P(chi^2 = 3, 8 \text{ d.f.}) > 0.9$	ns
GOE9	$P(chi^2 = 1.5, 5 \text{ d.f.}) > 0.9$	ns
GOE12	$P(chi^2 = 2.5, 5 \text{ d.f }) > 0.7$	ns
'Random'	$P(chi^2 = 9.8, 10 \text{ d.f.}) > 0.4$	ns

ns = not significant.

This shows that the numbers and distribution of all variants fit a random distribution. To test whether all types of variants are equally represented in the flanking regions, the cumulated totals of the 14 variants from all the flanking sequences were tested against the expected values.

Table 5.5 Observed and expected numbers of the GATA variants in the flanking regions.

	Observed				Expected*			
	Nucleotide replaced				Nucleotide replaced			
	G	A	T	A	G	A	T	A
Replaced by:								
A (0.31)	33	-	16	-	26.7	-	16.3	-
C (0.19)	12	6	14	8	16.3	10.0	10.0	10.0
G (0.19)	-	10	6	13	-	10.0	10.0	10.0
T (0.31)	22	9	-	20	26.7	16.3	-	16.3
deletion	-	25	41	-	-	42.7	36.4	-

total length analysed = 2889 nucleotides.

* Expected values were calculated by multiplying the total length analysed by the relative proportion of each nucleotide (in brackets above) in each variant sequence.

The observed results do not diverge significantly from those expected from a random distribution of nucleotide sequences ($P(chi^2 = 21.5, 13 d.f.)>0.05$; or $P(chi^2 = 13.6, 11 d.f.)>0.2$

if only the substitution variants are considered), taking into account the composition of the flanking regions.

5.6 Analysis of GATA variants within the GOE element

5.6.1 Determining a consensus sequence.

Although the GATA unit predominates, the relative position and/or type of some, or all, of the GATA variants may also be conserved between the GOE elements. In other words, can a consensus sequence be arrived at? The GOE elements are of different lengths and there is no *single* point common to all that appears to have strong claims for being a reasonable reference point. An alignment of sequences, as they are, is not possible. The diagram in figure 5.1 does not reveal an order of GATA variants that is consistent between the GOE elements. One can argue, though, that the occurrence of deletions within some members of an original GOE element would not only account for the differences in length, but would also alter the overall order of variants within the GOE elements suffering the deletion(s). On the proposal that the variants themselves are integral parts of GOE, each possible variant was given a code letter (these are explained in the legend to figure 5.5). The GOE elements are described in these terms in figure 5.5(a) and the order and type of the variants only is thus more readily visualised.

As the GOE elements are aligned in the figure (at their

5' ends), only in seven positions do two or more variants share the same position. This does not take into account the possibility that the GOE elements can be aligned differently, or that deletions may have occurred.

Figure 5.5(b) shows the order of the GATA units only. There is no common order shared by all the GOE elements. Only between GOE4 and GOE12 can two or more variants be found in the same order (an example is underlined in the figure). Even here, one finds that contiguous variants in the one GOE element are separated by GATA units in the other. Thus even if these two GOE elements were derived from a common sequence that incorporated some of the variants in their present order, deletions (or insertions) of GATA units would need to have occurred to produce the sequences as they are at present. In other words, the number of changes (a change here means single nucleotide substitutions, and deletions and insertions of any size) required to convert one GOE element to another is effectively equivalent to the sum of variants in each, regardless of whether some variants at particular positions are shared or not. The table below sums the equivalent changes needed to derive all the GOE elements from each other and from a contiguous poly(GATA) stretch. Three steps are needed to calculate the summed difference (e.g. for going from GOE5 to GOE6).

i) at least one insertion of a stretch of poly(GATA) to equalise lengths = 1 change.

ii) convert variants e,c in GOE5 to GATA = 2 changes.

iii) convert GATAs to l,l,w,y,z variants for GOE6 (w and z each require two changes) = 7 changes.

Table 5.5 Changes required to convert one GOE to another.

Original sequence	Final sequence					
	GOE5	GOE6	GOE9	GOE12	GOE4	Sum
GOE5	-	10	9	10	50	79
GOE6	10	-	13	14	56	103
GOE9	9	13	-	14	52	88
GOE12	10	14	14	-	55	93
GOE4	50	56	52	55	-	213
poly (GATA)	3	8	7	8	48	74

The poly(GATA) sequence therefore can be used to derive all the GOE elements most parsimoniously, followed by GOE5 which is itself predominantly GATA.

5.6.2 Substitutions of the poly(GATA) sequence.

Because poly(GATA) will derive all the GOE elements most parsimoniously, it is likely to be the closest to the ancestral sequence for all GOE elements. This cannot be proven, but would be further supported if the sequencing of all the GOE elements in the *D. melanogaster* genome showed that poly(GATA) is still the most parsimonious sequence. On the assumption that it is the ancestral sequence, then the forms of the present GOE elements could have arisen by the accumulation of base substitutions, deletions and insertions. The null hypothesis is that these changes accumulated in a random manner. Such a hypothesis can be tested in the following three ways.

a) There is an equal probability for substitution by each of the four nucleotides.

b) Each of the four positions in the GATA unit is equally likely to suffer a substitution.

c) Substitutions can occur equally throughout the poly(GATA) tract.

5.6.2a *Substitutions by the four nucleotides, A, C, G and T.*

If all nucleotides have an equal chance of substituting into a poly(GATA) sequence, the expected numbers are calculated as follows. In such a sequence, all substitutions with a C will be detected. Three quarters only of the substitutions with G or T will be detected, because one

quarter of the original sequence is already composed of these nucleotides. By the same argument, only half of the possible A substitutions can be detected. If n substitutions occur for each nucleotide, then the total (assuming only one substitution occurs at each site) is $4n$. The detectable substitutions, however, will be:

$0.5n$ (for A) + n (for C) + $0.75n$ (for G) + $0.75n$ (for T) = $3n$.

The apparent substitutions occurring in all the GOE elements can be summed, because one is asking whether all substitutions are equally possible in all GOE elements, not whether the degree of substitution is the same for all.

The total substitutions observed ($3n$) equals 58. The observed and expected substitutions are summarised in the table below.

Table 5.6 Substitutions into the poly(GATA) sequence.

Sequence	Length	Substitution				Total	Del	Ins	% changes per unit length
		A	C	G	T				
GOE4	424	5	5	10	20	39	6	0	9.2
GOE5	92	0	2	0	0	2	0	0	2.2
GOE6	192	0	1	0	3	4	1	1	2.1
GOE9	172	1	1	4	0	6	0	0	3.5
GOE12	260	0	0	1	6	7	0	0	2.7
Total changes		6	9	15	29	58	7	1	
Expected changes		9.6	19.3	14.5	14.5	58			

The observed substitutions deviate significantly from the expected values ($P(chi^2 = 26.9, 3\ d.f.) < 0.01$). The table shows that the greatest discrepancy lies in the T substitutions (more than twice that expected) and in the C substitutions (where there are half of what would be expected). If the contribution of GOE4 is excluded however, then the fit agrees with the expected totals ($P(chi^2 = 6.1, 3\ d.f.) > 0.05$).

5.6.2b *Substitutions into the four positions of the GATA unit.*

Is there a position in the GATA unit that is preferrably substituted? The numbers of substitutions at the four positions in all the GATA variants are listed in the table below, for all nucleotides and for T nucleotides only.

Table 5.7 Substitutions into the GATA unit

Sequence	Substitutions by all nucleotides				Substitutions by T nucleotides only			
	G	A	T	A	G	A	T	A
GOE5	1	0	1	0	0	0	-	0
GOE6	0	0	0	3	0	0	-	3
GOE9	2	0	3	1	0	0	-	0
GOE12	4	1	0	2	4	1	-	1
GOE4	16	9	6	9	14	4	-	2

If substitutions by T are excluded, the observed data fit well with the hypothesis that all four positions of the GATA unit are

equally likely to be substituted ($P(chi^2 = 3.8, 3\ d.f.)>0.2$). Of all the T substitutions (see table 5.6), half are at the G position of GOE4. If the GOE4 figures are adjusted so that T substitutions are equivalent in all three positions and all GOE elements and all positions are combined, the fit agrees well with an equal substitution at all four postions of the GATA unit ($P(chi^2 = 3.3, 3\ d.f.)>0.2$. Therefore, except for the predominance of TATA variants in GOE4, there is no statistically significant deviation from a random substitution at all positions in the GATA unit.

5.6.2c *Distribution of substitutions.*

The third aspect of the variants in GOE elements to be tested is their distribution along the sequences. A model of random substitution, deletion and insertion events to explain the form of GOE elements as they now are would suppose that the distribution of changes from a poly(GATA) sequence is essentially a random one. Such changes would be distributed so that the numbers (E) of GATA sequences of length, r nucleotides, that are uninterrupted by changes agree with the formula:

$$E = npq^r \quad \text{(Brown and Clegg, 1983)}$$

(where n is the total number of changes observed, p is the proportion of changes in the total length of sequence examined, and $q = 1 - p$.)

The combined results for all the GOE elements are presented in the table overleaf (Table 5.8). The observed distribution is

Table 5.8 Distribution of substitutions

Run length (r)	Observed*	Expected*
0	7	
1	6	8.56
2	1	
3	7	7.49
4	1	
5	4	6.56
6	5	
7	2	5.75
8	2	
9	1	5.04
10	4	
11	7	
12	1	6.42
13	1	
14	0	
15	1	5.26
16	1	
19	0	5.55
20	1	
21	0	
22	0	
23	0	
24	0	
25	0	
26	3	
27	2	
28	2	
29	0	
30	0	5.387

$n = 69$
$p = 0.064$
$q = 0.936$

$E = npq^r$

* Results are grouped so that expected values will exceed 5.0.

not significantly different from a random distribution of changes ($P(chi^2 = 13.2, 8\ d.f.)>0.1$. There appears to be an excess of changes lying adjacent to each other or separated by one base. Again most of the examples are from the GOE4 element.

In summary, the types and distribution of GATA variants discussed here are such that they do not disprove the null hypothesis that the GOE elements could have arisen by the random accumulation of point mutations into poly(GATA) sequences.

5.7 The TATA variants in the GOE4 element

There are 14 TATA units in the GOE4 sequence. Half of these lie in the central third of the sequence. A-T base pairs have lower thermal energies than G-C base pairs, so that a sequence rich in A and T nucleotides is less likely to maintain a duplex form than one with equal proportions of each. The 14 TATA units would increase the A-T content of GOE4 from 75% to about 77%. This is unlikely to affect the overall stability of the GOE4 sequence by a significant amount. TATAs are absent in GOE elements 5, 6 and 9. They are also virtually absent from the GOE elements of other organisms. For instance, in the mouse sequence, GOE(Mouse 1), 20% of the sequence is composed of GACA units and 10% of CATA units, while in the snake sequence, GOE(Snake), 27% is composed of GACA. TATAs make up only 1-2% in these sequences. Thus there is unlikely to be a universal role

for TATA in GOE elements.

Unequal crossovers, either between sister chromatids or between homologous chromosomes (Smith, 1976), could account for some of the apparent extra TATA units in the GOE4 element. Whenever a single TATA unit arises (by substitution at the G position), a series of crossover events can increase the number of TATAs without requiring further substitution events.

e.g. Strand 1 GATATATAGATA → GATATATATATAGATA
Strand 2 GATATATAGATA GATAGATA

and Strand 1 GATATATATATAGATA → GATATATATATATAGATA
Strand 2 GATATATATATAGATA GATATATATAGATA

There are two sets of sequences similar to that in strand 1 of the second example in GOE4, occupying positions 526 to 539 and 576 to 589. This could account for two TATAs without having to postulate T substitutions and for two of the double deletion events that were also postulated. The figures for the number of substitutions into each of the four positions of the GATA unit can therefore be adjusted to exclude the contribution of two of the TATA variants in the GOE4 element (see table 5.7). With this adjustment, the observed distribution does not deviate significantly from a model in which each of the four positions of the GATA unit is equally likely to be substituted ($P(chi^2 = 5.8, 3 d.f.)>0.1$).

Such unequal crossover events have also been postulated by Brown and Piechaczyk (1983) to account for the different numbers of contiguous CAAA units in two copies of the mouse MIF-1 family of repeated sequences (Brown and Dover, 1981).

5.8 Comparison of the GOE6 element from two *D. melanogaster* strains

The comparison of GOE elements within a genome suggests that they are all derived from a poly(GATA) sequence, and have accumulated base substitutions in essentially a random manner. There are fewer deletions and insertions than substitutions (8 as compared to 58 substitutions) and these are limited to GOE4 and GOE6. This would suggest that there has been little if any selective constraint on the types of changes that GOE elements can accumulate. The sequences of the GOE6 element from two fly strains (Canton S and FF) were also compared to see to what extent an individual copy is allowed to change.

The only differences going from the Canton S to the FF copy of GOE6 are, i) the addition of a GATA unit into the long GATA stretch covering positions 277 to 412, ii) the loss of three T's from positions 130 to 143 and iii) the replacement of a T with a G at position 129. Although all these changes can be treated formally as single base substitutions, they can also be regarded as deletions or insertions and could arise, through fewer steps, by unequal crossover events. These are illustrated overleaf.

i) - GATAGATAGATA - - GATAGATAGATAGATA -
X →
- GATAGATAGATA - - GATAGATA -

ii) - TTTGTTTTTTTTT - - TTTTGTTTTTTTTTTTT -
X →
- TTTTGTTTTTTTTT - - TTTTGTTTTTT -

iii) - TTTGTTTTTTTTT - - TTTGGTTTTTTTTT -
X →
- TTTGTTTTTTTTT - - TTTTTTTTTTTT -

All the differences between these two copies can be accounted for by such unequal crossovers. However, the examples in Canton S and FF cannot represent the respective reciprocal products of a single crossover event. Proof that such crossovers do occur can theoretically be obtained by comparing sequences of GOE6 from a founding generation of, say Canton S, to GOE6 sequences from later generations. This principle was used to demonstrate that unequal crossovers can occur in the *bobbed* locus in *Drosophila melanogaster* and so vary the numbers of ribosomal cistrons (Schalet, 1969). The advantage here was that the effect was detectable phenotypically. No phenotypic effect is known for GOE6 and crossover events would need to be determined at the sequence level. As the occurrence of detectable unequal crossover events between the tandemly arranged ribosomal cistrons is about 1 in 3000 generations (Frankham et al., 1980 and Coen et al., 1982), probably several thousands of lines would need to be sampled before one could argue that unequal crossovers do not occur in the GOE6 element. An alternative experiment would be to

sample GOE6 copies from a range of wild populations. This is essentially the same as the experiment suggested above, except that the sequence of the starting population cannot be sampled, and the original GOE6 element would be unknown. The advantage of using GOE6 for such an experiment is that it resides on the X chromosome, which is easier to isolate genetically than the autosomes.

It is interesting that no changes observed in the two GOE6 elements can be accounted for only by single base substitutions. Either the GOE6 elements are selected against accumulating changes or the time interval since the two populations separated is too short for these to have arisen. The former possibility is unlikely given the apparent random accumulation of changes already demonstrated.

5.9 Possible translation of the GOE elements

The 800bp Sau 3AI fragment from mouse reported in Epplen et al. (1983a) contained an open reading frame (ORF) throughout its length. The two published *Drosophila* GOE elements (Singh et al., 1984) also contain open reading frames. A search was therefore made of the GOE elements sequenced here, to see if similar ORFs were present.

Since GATA strands contain stop codons (TAG) every 12 bases, the complement strands were analysed. A poly(TATC) sequence cannot contain stop codons and will inevitably produce an ORF

that occupies its entire length. In contrast, a randomly assorted sequence would have, on average, a stop codon every 20 codons (or every 60 nucleotides). Therefore, the presence of a long (>20 codons) ORF in a GOE element is not as significant as the presence of a similarly sized ORF in a random sequence. It is important to see how far the ORFs extend beyond the GOE and into the flanking regions before supposing that GOE elements are likely to be translated because they contain long ORFs.

The sequences complementary to those presented in figure 4.2 were translated with the aid of a computer from positions 1, 2 and 3 and the longest open reading frames for each sequence identified.

To identify the most likely transcripts, the longest ORF that spans each GOE was first identified. In the case of GOE4, all three reading frames encountered one or more stop codons within the GOE element. Only the ORF that corresponded to the one published by Singh et al., (1984) was considered. Where two or three of the possible ORFs for a particular GOE were of similar lengths, only those that contained a methionine residue (start codon) upstream of the GOE element and within the ORF were considered. In the event, only for GOE5, GOE6 and GOE9 could such start codons be found. For both GOE4 and GOE12 the methionines nearest to the GOE region were separated from it by stop codons. Though an intervening intron sequence could obviate this problem, the only identifying signals (GT at the 5' end and AG at the 3' end) would be so common in any sequence as to produce too many potential open reading frames.

Translated sequences should also be flanked i) by the consensus sequence for the RNA polymerase binding site (the TATA box which has the sequence: $TAT^{TA}_{AT}A^{T}_{A}$) lying 40 to 80 bases upstream from the start codon and ii) by a polyadenylation signal sequence (AATAAA, AATTAAA or ATTAAA) lying downstream from the stop codon (though not all messenger or polysomal sequences possess poly(A) tails).

Of GOE5, GOE6 and GOE9, only GOE9 possesses sequences similar to TATA boxes lying upstream from the start codon. These lie at positions 320-330 and 350-360 on the GATA strand. The sequenced region of GOE9 does not extend beyond the proposed translated region, so a polyadenylation site cannot be searched for. Only GOE4 has a potential polydanylation signal sequence, (at positions 170-180 of the GATA strand) which would place it 300bp downstream from the stop codon.

The upstream regions of GOE4 and GOE12 are quite AT-rich, and several TATA-like sequences are found. These are not surrounded by GC-rich regions, which is another requisite for RNA polymerase binding sites. Also, because no start codon can be found for GOE4 or GOE12, without proposing the presence of introns, this makes identification of feasible ORFs for all GOE elements except GOE9 very indefinite.

The best evidence would be the existence of a transcript *in vivo*. This has been found for mouse (e.g. Epplen et al., 1982 and Singh et al., 1984) and is discussed in more detail in the next chapter. Preliminary experiments probing Northern blots containing a range of *Drosophila* RNAs from different stages of

the life-cycle with both strands of GOE6 did not detect any discrete bands (M. Healy, pers. comm.). This suggests that GOE elements in *Drosophila* are not transcribed to an appreciable extent. Even if a transcript had been identified, this need not have corresponded to the particular GOE element used as a probe. The corresponding cDNA clone would need to be sequenced to identify which GOE was being transcribed.

The translation product of a continuous TATC sequence would be a repeat of the tetramer, Tyrosine-Leucine-Serine-Isoleucine. This puts a polar amino acid (serine) in the midst of three hydrophobic amino acids. Such a peptide would require a periodic structure in which the serine residues are localised in one area and the hydrophobic residues in another, so that the molecule as a whole is divided into hydrophilic and hydrophobic domains. The bulk of such a peptide would have to reside in a hydrophobic environment, such as a cellular membrane.

5.10 Analysis of the non-*Drosophila* GOE elements

The general characteristics that are revealed by the *Drosophila* GOE elements are present also in the published non-*Drosophila* GOE elements. That is, they all contain a number of localised and contiguous GATA tetranucleotides which are generally separated by sequences that can be converted to GATA units by a few nucleotide substitutions, deletions or insertions. Beyond the GATA-rich regions, there is no immediately apparent homology between the GOE elements.

A more detailed analysis similar to that performed on the *Drosophila* GOE elements was therefore carried out on these non-*Drosophila* sequences, to see

a) whether the sequences flanking the GOE elements are homologous to each other,

b) if they are not homologous to each other, whether they are formally equivalent to randomly generated sequences and

c) whether the GOE elements could have arisen by the accumulation of random mutations into a poly(GATA) sequence.

5.11 The limits of the non-*Drosophila* GOE elements

The non-*Drosophila* GOE elements, whose sequences have been published, will be referred to as follows:

a) GOE(Mouse 1) element = clone *pmc14* of Epplen et al., (1983a,b).

b) GOE(Mouse 2) element = clone *M3.1* of Singh et al., (1984).

c) GOE(Rat) element = clone *pAF4* of Alonso et al., (1983).

d) GOE(Snake) element = clone *pErs5* of Epplen et al., (1982).

The non-*Drosophila* GOE sequences are presented schematically in figure 5.1, in the same way as the *Drosophila* sequences (described in section 5.1). As in the case of the *Drosophila* sequences, these are also of different lengths and have different distributions of GATA units. To provide a more accurate definition for these GOE elements, the last pair of GATA units was taken as the 3' end of each GATA region. Similarly, the first pair of GATA units was initially considered to be the 5' end. However, the first pair of GATA units in the GOE(Mouse 1) and GOE(Mouse 2) sequences are preceded by variants of GATA and by a single GATA unit, and this GATA unit was taken as the 5' limit of these two mouse sequences.

The GOE(Rat) element may well extend beyond the 3' end of the sequenced region. It may also be described as two closely apposed GOE elements, since the sequence between the two main blocks of GATAs are not readily convertible to GATA units.

The positions of these limits are indicated in figure 5.1

and the GATA regions themselves are printed in upper case letters in figure 5.6.

5.12 Search for sequence similarities beyond the GATA regions

5.12.1 Comparison of the sequences immediately adjacent to the GATA regions.

The 15 nucleotides immediately preceding and following the GATA regions are listed in figure 5.7. Unlike the case for the *Drosophila* GOE elements, there is no AT-rich 5' flanking sequence and the 3' region does not consist of GATA variants.

To search for any significant sequence similarity between the flanking regions on a broader scale, the two computer-assisted methods described in sections 5.2.2 and 5.2.3 were applied to these non-*Drosophila* GOE elements.

5.12.2 Comparison of the sequences beyond the GATA regions by the Dayhoff alignment method.

The 200 nucleotides on either side of the GATA regions for each GOE element were aligned using the Dayhoff method, and the resulting alignment scores are presented in Table 5.9. In three cases the alignment scores between the 3' flanking regions exceed a value of 2.0, though the two mouse sequences do not show any significant similarity.

Table 5.9 Alignment scores for the flanking sequences of the non-*Drosophila* GATA regions

5' flanking region

	GOE(Mouse1)	GOE(Mouse2)	GOE(Snake)
GOE(Mouse 1)	-	0.43ns	-1.56ns
GOE(Mouse 2)	0.43ns	-	1.66ns
GOE(Snake)	-1.56ns	1.66ns	-
GOE4	-0.59ns	-1.42ns	-1.46ns
GOE5	1.67ns	-0.91ns	-0.74ns
GOE6	-2.50*	-3.12*	-0.09ns
GOE12	0.18ns	0.12ns	0.45ns

3' flanking region

	GOE(Mouse1)	GOE(Mouse2)	GOE(Snake)
GOE(Mouse 1)	-	-1.25ns	1.44ns
GOE(Mouse 2)	-1.25ns	-	0.10ns
GOE(Snake)	1.44ns	0.10ns	-
GOE4	-2.48*	-1.05ns	-2.92*
GOE5	-0.40ns	2.97*	2.20*
GOE6	0.02ns	2.55*	0.21ns
GOE9	-0.18ns	1.74ns	0.17ns
GOE12	0.14ns	0.67ns	0.28ns

ns = not significant.

* = <5% probability of the alignment occurring by chance.

5.12.3 Comparison of the sequences beyond the GATA regions by the dot matrix diagram method.

The 200 nucleotides immediately adjacent to the GOE elements were compared by the dot matrix method. No matches of 10 or more base pairs were detected. The total numbers of 4bp matches for each comparison are presented in table 5.10. In three cases*, the observed number of matches significantly exceeds the expected mean (= 142.6 ±17.2). All of these examples are from the comparisons between the non-*Drosophila* and the *Drosophila* flanking regions and not between the mouse GOE elements' flanking regions. Furthermore, none of the pairwise comparisons that were shown by this method to be significant correspond to those that were shown to have significant alignments (previous section).

The comparison of the flanking regions of the *Drosophila* and non-*Drosophila* GOE elements shows there to be no consistent and significant similarity between them. This confirms that the GATA regions, as they are defined here, must be the only common sequences between the various GOE-containing sequences and are therefore a sufficient definition for the GOE element itself.

* The expected number of matches is obtained by assuming that all the sequences tested have the same (A+T) content (= 0.6). Because the 3' flanking regions of GOE12 and of the non-Drosophila all have (A+T) contents greater than 0.6, they will match up more frequently anyway. The significantly large number of matches obtained need not imply that these sequences have a structure in common. Conversely, the 5' flanking region of GOE6 has a small (A+T) content (0.45) and will therefore give fewer matches with the corresponding regions of the non-Drosophila GOE sequences than would be expected.

Table 5.10 Number of 4bp matches between the flanking sequences of the non-*Drosophila* GATA regions

5' FLANKING REGION

	GOE(Mouse1)	GOE(Mouse2)	GOE(Snake)
GOE(Mouse 1)	-	151ns	116ns
GOE(Mouse 2)	151ns	-	111ns
GOE(Snake)	116ns	111ns	-
GOE4	109*	120ns	122ns
GOE5	142ns	143ns	168ns
GOE6	140ns	120ns	162ns
GOE12	118ns	122ns	138ns

3' FLANKING REGION

	GOE(Mouse1)	GOE(Mouse2)	GOE(Snake)
GOE(Mouse 1)	-	114ns	116ns
GOE(Mouse 2)	114ns	-	118ns
GOE(Snake)	116ns	118ns	-
GOE4	141ns	127ns	149ns
GOE5	124ns	150ns	122ns
GOE6	109*	106*	101*
GOE9	122ns	140ns	156ns
GOE12	184*	191*	238*

Expected mean no. of 4bp matches = 142.6 ±17.2.

ns = not significant. * = significant.

5.13 Analysis of the GATA variants in the sequences flanking the GATA regions of the non-*Drosophila* GOE elements

5.13.1 Numbers and types of GATA variants.

The numbers and types of GATA variants in the GATA and flanking regions are presented in table 5.11 (overleaf). The total values for the 5' and 3' flanking regions are listed in the table following, together with the values that are expected on the basis of a random assortment of nucleotides (table 5.12).

Table 5.11 GATA variants in the non-*Drosophila* GOE-containing sequences

GOE(MOUSE 1) SEQUENCE

Region	Replacement	Nucleotide replaced				
		G	A	T	A	
	A	2	0	9	0	11
5' flanking	C	5	3	13	3	24
region	G	0	4	8	1	13
n=1271	T	7	3	0	0	10
	total replacments	14	10	30	4	58
	deletions	0	19	26	0	45
	total changes	14	29	56	4	103

Region	Replacement	Nucleotide replaced				
		G	A	T	A	
	A	0	0	2	0	4
GOE region	C	9	0	33	0	42
n=676	G	0	4	5	3	12
	T	2	0	0	0	2
	total replacements	11	4	40	3	58
	deletions	0	5	2	0	7
	total changes	11	9	42	3	65

Region	Replacement	Nucleotide replaced				
		G	A	T	A	
	A	0	0	4	0	4
3' flanking	C	2	4	1	1	8
region	G	0	1	2	1	4
n=455	T	2	2	0	1	5
	total replacements	4	7	7	3	21
	deletions	0	6	9	0	15
	total changes	4	13	16	3	36

Table 5.11 (contd.)

GOE(MOUSE 2) SEQUENCE

Region	Replacement	Nucleotide replaced				
		G	A	T	A	
	A	1	0	2	0	3
5' flanking	C	1	0	1	1	3
region	G	0	0	2	1	3
n=144	T	0	0	0	0	0
	total replacements	2	0	5	2	9
	deletions	0	6	5	0	11
	total changes	2	6	10	2	20

Region	Replacement	Nucleotide replaced				
		G	A	T	A	
	A	0	0	0	0	0
GOE region	C	1	0	1	0	2
n=98	G	0	0	0	0	0
	T	0	0	0	0	0
	total replacements	1	0	1	0	2
	deletions	0	0	1	0	1
	total changes	1	0	2	0	3

Region	Replacement	Nucleotide replaced				
		G	A	T	A	
	A	1	0	0	0	1
3' flanking	C	0	1	3	0	4
region	G	0	1	1	1	3
n=177	T	2	0	0	1	3
	total replacements	3	2	4	2	11
	deletions	0	7	2	0	9
	total changes	3	9	6	2	20

Table 5.11 (contd.)

GOE(RAT) SEQUENCE

Region	Replacement	Nucleotide replaced				
		G	A	T	A	
	A	1	0	1	0	2
5' flanking	C	0	1	0	0	1
region	G	0	0	0	1	1
n=32	T	0	0	0	0	0
	total replacements	1	1	1	1	4
	deletions	0	0	1	0	1
	total changes	1	1	2	1	5

Region	Replacement	Nucleotide replaced				
		G	A	T	A	
	A	2	0	1	0	3
GOE region	C	0	0	2	1	3
n=218	G	0	2	0	2	4
	T	1	2	0	1	4
	total replacements	3	4	3	4	14
	deletions	0	2	1	0	3
	total changes	3	6	4	4	17

Table 5.11 (contd.)

GOE(SNAKE) SEQUENCE

Region	Replacement	Nucleotide replaced				
		G	A	T	A	
5' flanking region n=1530	A	10	9	9	0	28
	C	7	4	4	3	18
	G	0	3	5	5	13
	T	6	7	0	9	22
	total replacements	23	23	18	17	81
	deletions	0	16	28	0	44
	total changes	23	39	46	17	125

Region	Replacement	Nucleotide replaced				
		G	A	T	A	
GOE region n=162	A	0	0	0	0	0
	C	0	1	12	0	13
	G	0	0	0	0	0
	T	1	0	0	0	1
	total replacements	1	1	12	0	14
	deletions	0	0	0	0	0
	total changes	1	1	12	0	14

Region	Replacement	Nucleotide replaced				
		G	A	T	A	
3' flanking region n=791	A	13	0	6	0	19
	C	4	3	3	1	11
	G	0	2	0	4	6
	T	9	7	0	3	19
	total replacements	26	12	9	8	55
	deletions	0	9	11	0	20
	total changes	26	21	20	8	75

Table 5.11 (contd.)

Totals for all non-*Drosophila* GOE elements

Region	Replacement	Nucleotide replaced				
		G	A	T	A	
5' flanking region n=2977	A	14	0	21	0	35
	C	13	8	18	7	46
	G	0	7	15	8	30
	T	13	10	0	9	32
	total replacements	40	25	54	24	143
	deletions	0	41	54	0	95
	total changes	40	66	108	24	138

Region	Replacement	Nucleotide replaced				
		G	A	T	A	
GOE region n=1154	A	2	0	3	0	5
	C	10	1	48	1	60
	G	0	6	5	5	16
	T	3	2	0	1	6
	total replacements	15	9	56	7	87
	deletions	0	9	6	0	15
	total changes	15	18	62	7	102

Region	Replacment	Nucleotide replaced				
		G	A	T	A	
3' flanking region n=1434	A	14	0	10	0	24
	C	6	8	9	2	25
	G	0	4	3	6	13
	T	12	9	0	5	26
	total replacements	32	21	22	13	88
	deletions	0	22	23	0	45
	total changes	32	43	45	13	133

Table 5.12 Observed and expected GATA variants in the flanking regions

5' FLANKING REGION

	Observed				Expected			
	G	A	T	A	G	A	T	A
Substituted by:								
A	14	-	21	-	32.0	-	32.0	-
C	13	8	18	7	14.5	10.0	14.0	10.0
G	-	7	15	8	-	15.0	21.0	15.0
T	13	10	-	9	22.8	15.6	-	15.6
deletion		41	54			46.9	55.6	

3' FLANKING REGION

	Observed				Expected			
	G	A	T	A	G	A	T	A
Substituted by:								
A	14	-	10	-	12.6	-	11.7	-
C	6	8	9	2	6.3	4.9	5.8	4.9
G	-	4	3	6	-	7.7	9.2	7.7
T	12	9	-	5	10.7	8.4	-	8.4
deletions		22	23			23.2	25.6	

Though the numbers and types of GATA variants in the 3' flanking region do not show a significant deviation from what would be expected ($P(chi^2 = 13.1, 13 \text{ d.f.}) > 0.4$), this is not the case in the 5' flanking region ($P(chi^2 = 35.4, 13 \text{ d.f.}) < 0.01$). In this latter region, there are fewer AATA, GAAA and TATA variants than would be expected. Why this should be so is not clear.

5.13.2 The distribution of GATA variants along the sequences flanking the GOE elements.

The distribution of variants along the flanking regions is essentially what would be expected (see section 6.6). That is, assuming that 1.48 variants would be expected in every 20 nucleotides of sequence, the flanking regions of the non-*Drosophila* GOE elements do not deviate significantly from this distribution. i.e.

GOE(Mouse 1)	$P(chi^2 = 16.5, 28\ d.f.) > 0.9$	ns
GOE(Mouse 2)	$P(chi^2 = 5.7, 5\ d.f.) > 0.3$	ns
GOE(Snake)	$P(chi^2 = 21.7, 37\ d.f.) > 0.9$	ns

ns = not significant.

(The GOE(Rat) element is not included, because there are only 40 nucleotides of sequence outside the GATA region).

5.14 Analysis of GATA variants within the GOE elements

Assuming that the non-*Drosophila* GOE elements, like the *Drosophila* ones, were also derived from poly(GATA) sequences, then the substitutions that would be required to produce the present sequences are listed below:

Table 5.13 Substitutions into a poly(GATA) sequence.

Sequence	Length	Substituted by:				Total	% changes per unit length
		A	C	G	T		
GOE(Mouse 1)	676	2	42	12	2	58	8.6
GOE(Mouse 2)	98	0	2	0	0	2	2.0
GOE(Rat)	217	3	3	4	4	13	5.9
GOE(Snake)	161	0	13	0	1	14	8.7
Total		5	60	16	6	87	
Expected total		15	29	22	22		

Table 5.14 Substitutions into the four positions of the GATA unit.

Sequence	Substitution into:			
	G	A	T	A
GOE(Mouse 1)	11	4	40	3
GOE(Mouse 2)	1	0	1	0
GOE(Rat)	2	4	3	4
GOE(Snake)	1	1	12	0
Total:	15	9	56	7
Expected Total:	22	22	22	22

Obviously, these results are not consistent with the proposition that these sequences were derived solely by random substitution into poly(GATA) sequences. Substitutions by a C nucleotide and substitutions into the third (T) position of the GATA unit are very common. In fact, GACA is the predominant variant in the GOE(Mouse 1) and GOE(Snake)

sequences. Ten of the twelve GACAs in GOE(Snake) are tandemly arranged at the 3' end of this sequence. On the other hand, the GACA variants are more evenly dispersed in the GOE(Mouse 1) sequence, though they all reside in sequences of the form:- $(GACA)_{2-3}TAT$ or $(GACA)_2T$.

However, neither tandem stretches of GACA nor $(GACA)_{2-3}T(AT)$ sequences are present in any of the other GOE elements. If they do serve some sequence dependent function, it is not universal in all the GOE elements. Some of these 'GACA' sequences could have arisen by mechanisms other than random substitution into a poly(GATA) sequence, for example by unequal crossover (Smith, 1976).

If the contribution of these GACA variants is excluded, then the substitutions into the GOE elements do not deviate significantly from the expected values (for substitutions by A,C G or T: $P(chi^2 = 4.3, 3 \text{ d.f.})>0.1$ and for substitutions into the GATA unit: $P(chi^2 = 5.4, 3 \text{ d.f.})>0.05$).

The distribution of substitutions was also analysed, in the same way as described in section 5.7.2. On a random basis, the expected numbers (E) of sequences of length (r) that are not interrupted by apparent substitutions is given by the formula:

$E = npq^r$ (n, p and q have been defined previously).

Table 5.15 overleaf combines the values for all the non-*Drosophila* GOE elements.

Table 5.15 Distribution of apparent substitutions into poly(GATA) sequences.

Run length (r)	Observed	Expected
0	29	26.5
1	31	22.6
2	22	19.1
3	46	16.3
4	6	13.8
5	8	11.8
6	5	10.0
7	3	8.5
8	2	7.2
9	2	6.1
10	3	5.2
11	2	
12	0	8.2
13	0	
14	1	5.9
15	0	
16	1	
17	0	5.9
18-22	4	5.2

$n = 177$
$p = 0.15$
$q = 0.85$

$E = npq^r$

The distribution of apparent substitutions is not equivalent to this model. It has already been pointed out that there is an apparent excess of GACA variants in the GOE(Mouse 1) and GOE(Snake) elements. Tandem stretches of *n* GACA variants will be interpreted here as equivalent to n-1 substitutions separated by a run length of 3 nucleotides and will account to some extent for the non-random distribution. Yet, even if the GACA variants in the GOE(Mouse 1) and GOE(Snake) sequences are excluded, the observed values still deviate significantly ($P(chi^2 = 35.4, 13\ d.f.)<0.05$).

5.15 Summary of the analysis of non-*Drosophila* GOE elements

In summary, the non-*Drosophila* GOE elements show similar properties to their *Drosophila* counterparts. Their flanking sequences are not similar to each other and are essentially composed of a random distribution of nucleotides. The (A+T) contents of these flanking sequences are also similar to the average (A+T) content for most vertebrate genomes.

There is no conserved arrangement of GATA variants within the GOE elements themselves, though the numbers and distributions of these variants is not consistent with their having arisen by the accumulation of random mutations into a poly(GATA) sequence. The GACA variant is prominent in two of the GOE elements but is virtually absent in the others, including the *Drosophila* sequences. Like the TATA variant in

the *Drosophila* GOE element, GOE4, some of these 'extra' GACA variants could have arisen by unequal crossover events, rather than by substitution into a GATA unit.

The four non-*Drosophila* GOE elements are from three different species and do not represent a major intragenomic survey as does the analysis of the *Drosophila* sequences. Without the analysis of a large portion of the mouse GOE element complement, for example, it would be unwise to impute the action of selective forces to explain the non-random nature of the apparent mutations that have generated the present form of the non-*Drosophila* GOE elements.

5.16 Overall summary of the analysis of the GOE elements

The *Drosophila* GOE elements are composed predominantly of tandem repetitions of the tetranucleotide, GATA. They are situated amongst sequences that are not related to each other. These flanking regions have (A + T) contents similar to that of the bulk of *Drosophila* genomic DNA and they have nucleotide distributions that would be expected from randomly generated sequences.

The structures of the GOE elements are dissimilar from those repetitive sequences discussed in chapter 1 (and illustrated in figure 1) and they are not flanked by the short direct repeats that are thought to indicate insertion of a mobile sequence into the genome. Such short direct repeats are usually of the same length but differ in sequence between copies. Therefore, the first and last GATAs of the GOE elements are not analogous to these.

The five GOE elements are of different lengths and have different numbers and types of non-GATA sequence. The non-GATA sequences can however be converted into GATAs by one or two substitutions, deletions or insertions of single nucleotides. The non-GATA sequences are termed GATA variants. All the GOE elements can be most parsimoniously derived from a poly(GATA) sequence. The numbers, types and distribution of the GATA variants is consistent with a model that they arose from poly(GATA) sequences by the accumulation of random substitution, deletion and insertion events.

The GOE6 copies from the Canton S and FF strains differ only by the addition of one GATA unit into the latter (or the loss of one unit from the Canton S copy). Changes in the number of GATA units are most easily envisaged as taking place by unequal crossover, rather than by single base substitution.

Without knowing the ancestral sequence (or sequences) one cannot know if there has been any selective constraint on the type of mutation in GATA, though the evidence here suggests that there has not.

Drosophila GOE elements are flanked by an AT-rich region at the 5' end, but this is not seen in the non-*Drosophila* GOE elements, and so cannot be a universal part of GOE. The non-*Drosophila* GOE elements also contain GATA variants, but some are tandemly repeated within the GOE element. For example, the CATA stretch at the beginning of the GOE(Mouse 1) element or the GACA stretch at the end of the GOE(Snake) element. Therefore, the only common aspect of the GOE element both between and within species is the localised distribution of tandemly repeated GATA units.

Figure 5.1

Diagram of the GOE sequences discussed in the text. Single blocks represent 4 nucleotide units. Fully shaded blocks correspond to GATA units, half shaded blocks to single nucleotide variants of GATA and unshaded regions represent all other sequences.

Numbers correspond to the positions that delimit the GATA regions of the *Drosophila* sequences as they are defined in the text (section 5.1).

All or parts of the published non-*Drosophila* GOE sequences are also presented.

Mouse1 = pmc14 sequence (positions 1162 - 2095) from Epplen et al., 1983b.

Mouse2 = mouse clone (positions 1 - 420) from Singh et al., 1984.

Rat = rat repetitive DNA clone (positions 1 - 261) from Alonso et al., 1983.

Snake = pErs5 sequence (positions 1210 - 1950) from Epplen et al., 1982.

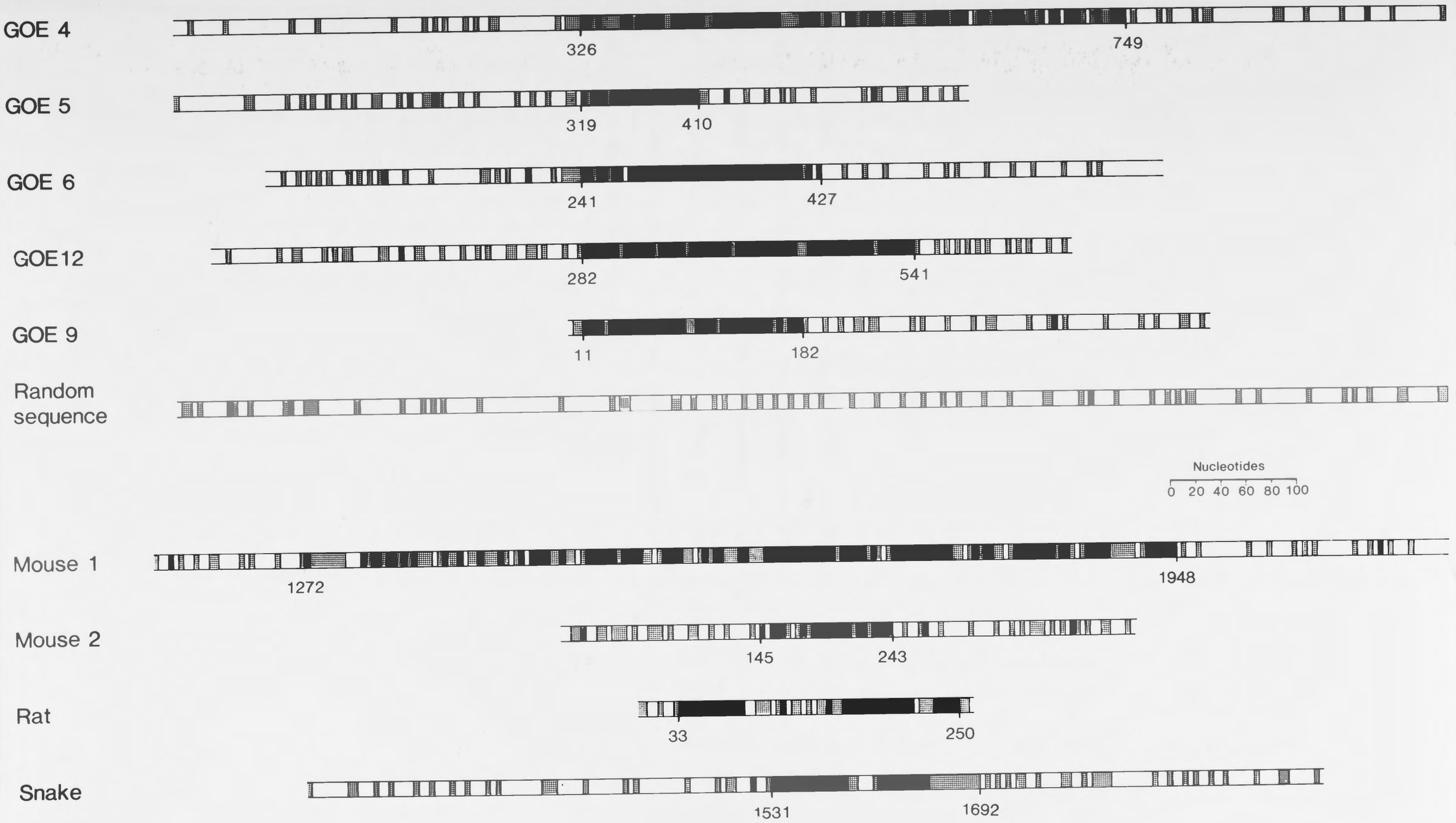

GOE 4
326
749
GOE 5
319
410
GOE 6
241
427
GOE12
282
541
GOE 9
11
182
Random sequence
Nucleotides
0 20 40 60 80 100
Mouse 1
1272
1948
Mouse 2
145
243
Rat
33
250
Snake
1531
1692

Figure 5.2

Alignment of the 20 nucleotides on the 5' side and 15 nucleotides on the 3' side of the GATA regions in the five *Drosophila* GOE sequences. Some sequences are adjusted to provide the greatest number of matches between the regions. Those parts where three or more nucleotides at a particular position are shared are boxed.

Sequence	5' region	GATA region	3' region
	A		
GOE4	GGATGATGAATATATATATATA	.....	GTTTGATTGCAACCT
GOE5	TTAGACGGCATACATATAT-	.....	TATAGATTTGACTAT
GOE6	TGAGAAATCATAAATAGA-A	.....	AATTGCACATGCTTC
GOE9	TATATATATA	.....	GATTCCTATGAATGA
GOE12	TCTGTTGATTTGTCAAAATA	.....	GATGGTCGACTACTT

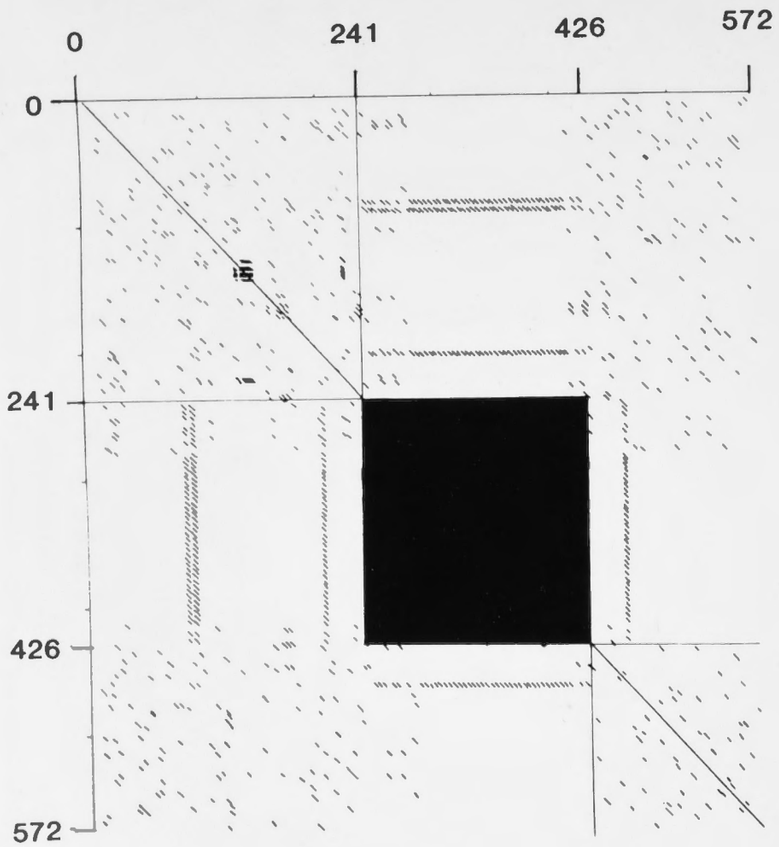
Figure 5.3

Dot matrix diagrams for three pairwise comparisons involving GOE6. The GOE6 sequence is arranged on each horizontal axis and GOE6, GOE5 and GOE12 arranged on the vertical axes. Numbers indicate the start and finish positions of the whole sequences tested, and the start and finish of the GATA regions, for each GOE sequence.

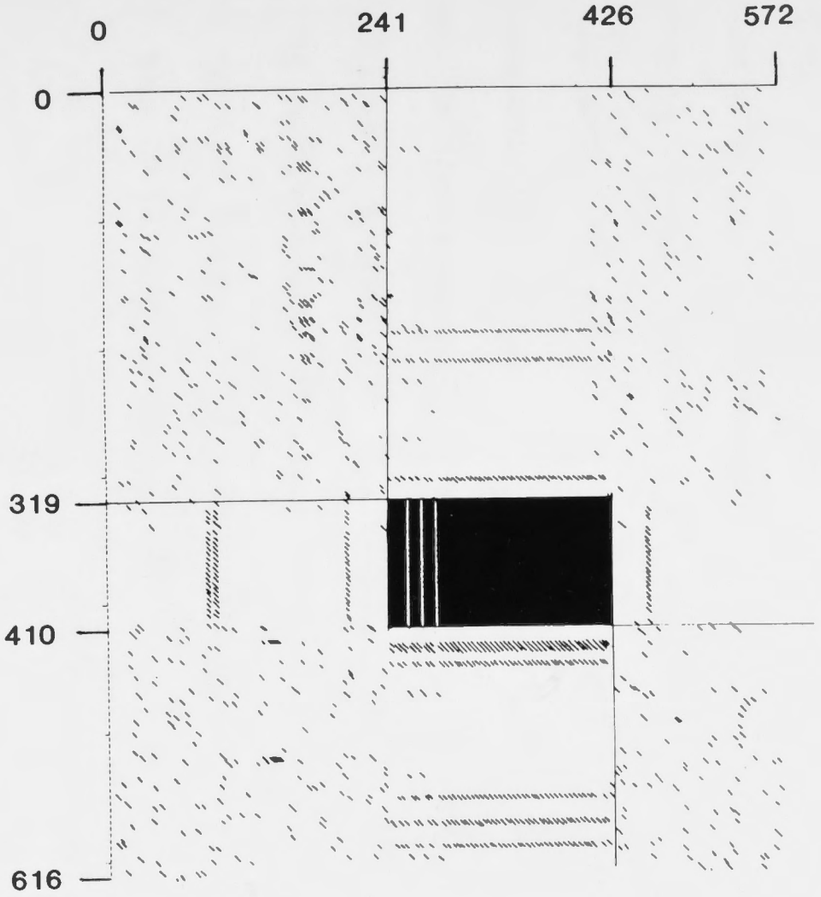
Each diagonal mark represents a match of at least 4 nucleotides between the two sequences being compared. Because matches are detected throughout when two GATA regions are compared, these are represented by fully shaded blocks in the figure.

DOT MATRIX DIAGRAMS OF GOE SEQUENCES

GOE6 vs GOE6



GOE6 vs GOE5



GOE6 vs GOE12

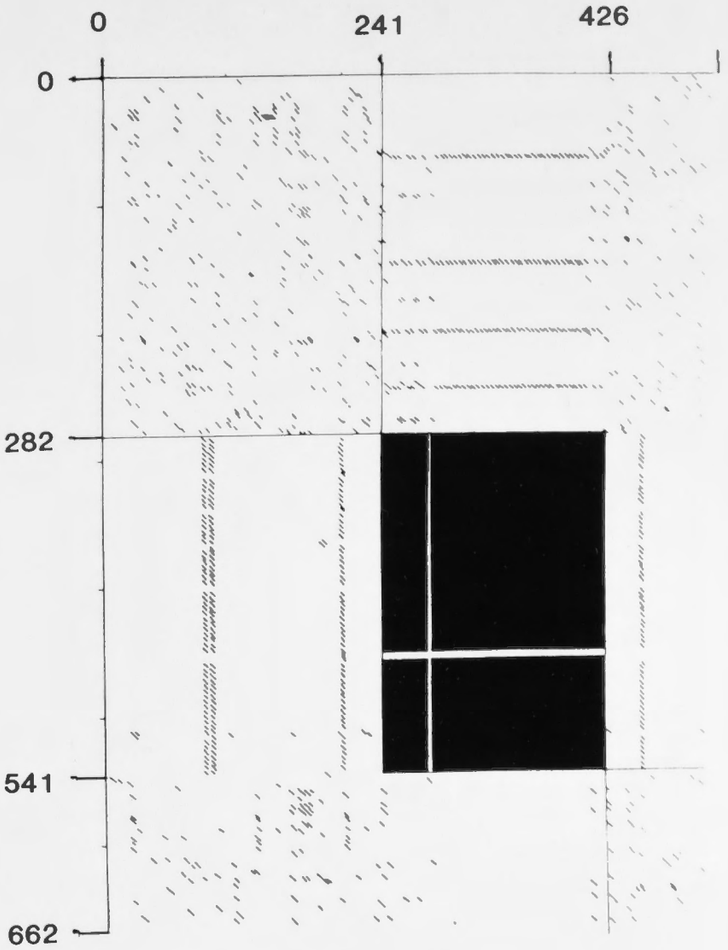
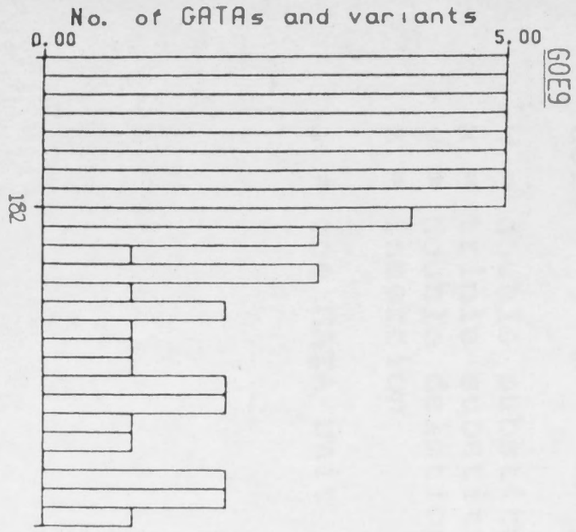


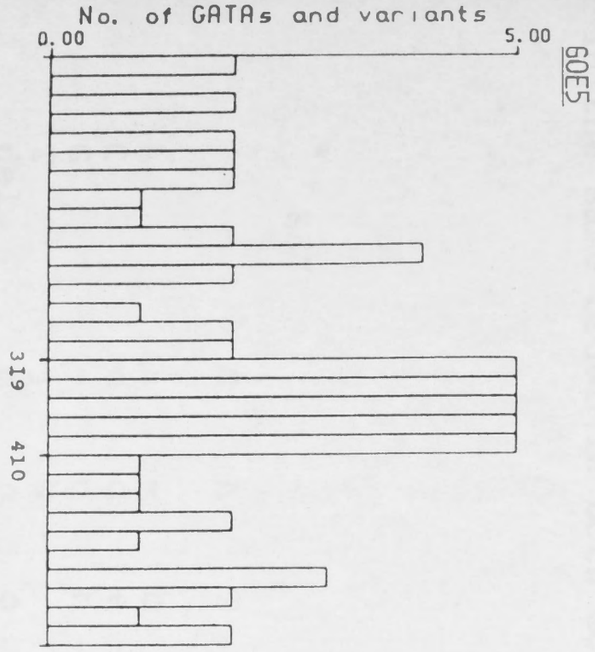
Figure 5.4

Histogram plots of the numbers of GATAs and GATA variants in blocks of 20bp. The five GOE sequences were divided into blocks of 20bp and the number of GATAs and GATA variants in each block counted (no more than five are possible). These are presented in the figure in form of histogram plots. The plot for a randomly generated 600bp sequence with equal proportions of nucleotides is also presented for comparison.

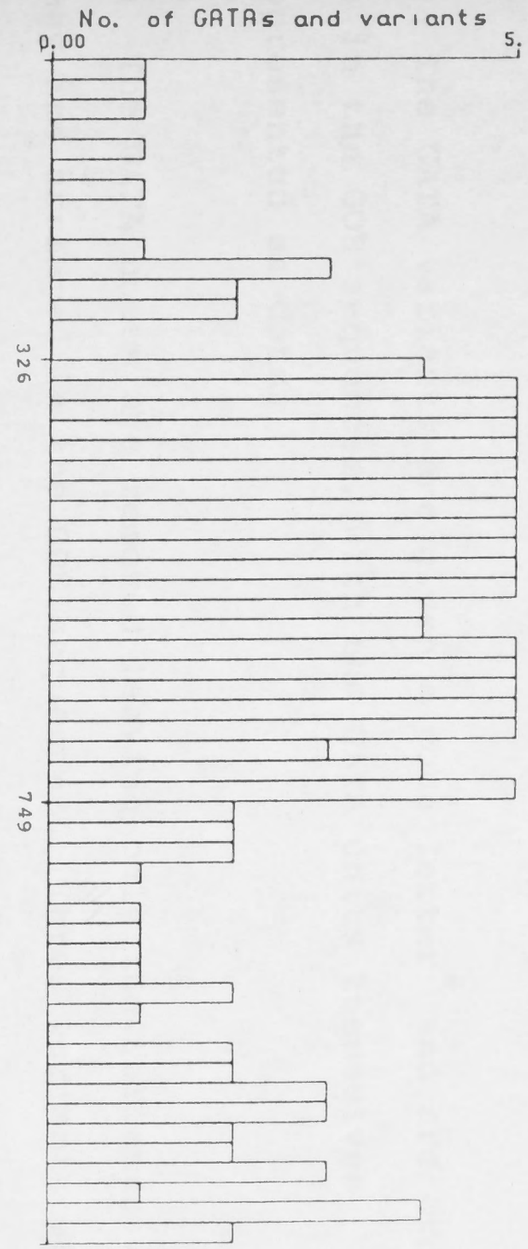
Numbers on the X-axes indicate the start and finish positions of the GATA regions.



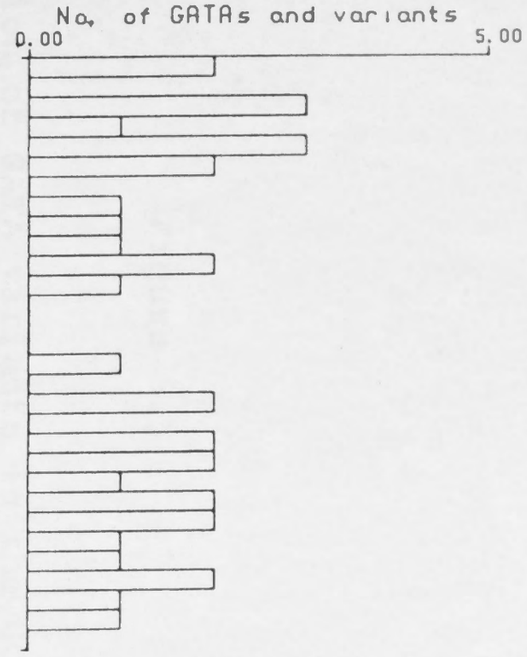
GOE9



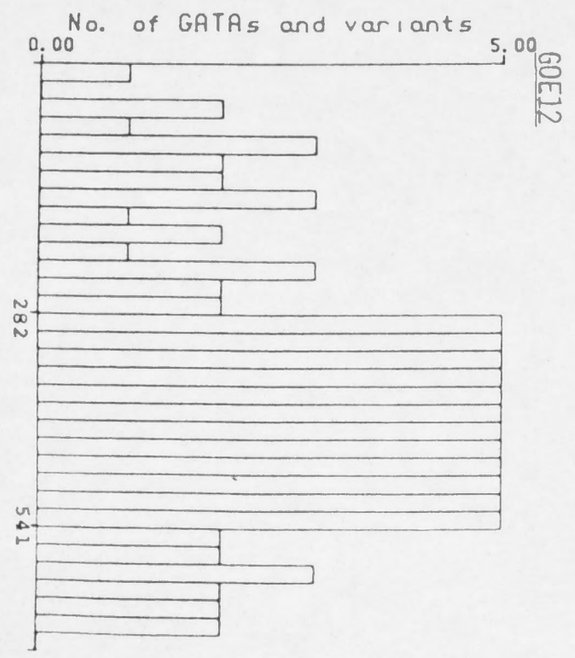
GOE5



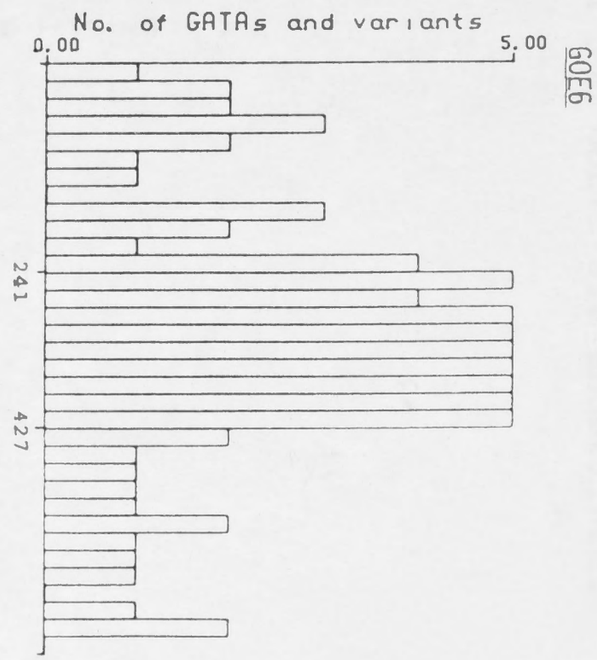
GOE4



RANDOM



GOE12



GOE6

Figure 5.5

Representation of GATA variants in the GOE sequences.

a) The GATA variants are given a code letter* and are ordered as in the GOE sequences, with the GATA units themselves represented as dots.

b) The GATA units are removed leaving only the variants as they are arranged in the GOE sequences. Three variants which are in the same order in both GOE4 and GOE12 are underlined.

* KEY

	G	A	T	A
A	a	-	b	-
C	c	d	e	f
G	-	g	h	i
T	j	k	-	l
del	m	n	o	p

w = double substitution
x = triple substitution
y = double deletion
z = insertion

. = one GATA unit

Figure 5.5

a)

```
GOE5    .e...c.................
GOE6    ..l..l..w.....................................y.z.
GOE9    ....h..............hh....c..........i.a...
GOE12   .......j......j.....i........k.............jj..............l.......
GOE4    ..j.ijik..g......w......ik.............jlme....j.jjy..g.j..w.j.jjy/
        l.g....i...gm..j......e.m.w..w..fxa.pk..
```

b)

```
GOE5    ec
GOE6    llwzy
GOE9    hhhcia
GOE12   jjikjjl
GOE4    jijikgwikjlmejjjygjwjjjylgigmjemwwfxapk
```

Figure 5.6

The non-*Drosophila* GOE sequences.

Parts of the non-*Drosophila* GOE sequences which have been published elsewhere are presented here. The GATA regions are shown (in upper case letters) along with 50 to 100 nucleotides of the flanking sequences.

GOE(Mouse 1) is from Epplen et al., (1982).

GOE(Mouse 2) is from Singh et al., (1984).

GOE(Rat) is from Alonso et al., (1984).

GOE(Snake) is from Epplen et al., (1981).

a) GOE(Mouse1)

```
      1210       1220       1230       1240       1250       1260       1270       1280       1290       1300
aaatgtagct caaggctgtc ctctgatctc tatatgttta ctatggcatg tatgctcccc acacgcagac aGATACATAC ATACATACAT ACATATATAC

      1310       1320       1330       1340       1350       1360       1370       1380       1390       1400
ATACACACAC ACAGACTGAT AGATGATAGA TACATAGATA GATAGATGAT AGGTAGATAG ATGAGTAGAC AGACATATGA TAGGTAGATG ATATAGACAG

      1410       1420       1430       1440       1450       1460       1470       1480       1490       1500
CATATGATAG GTAGATGGAT GATAGACAGA CATATGATAG GTAGATAGAT AGATAGATAG ACAGACAGAT AGATGAAAGA CAGACATATG ATAGATAGAT

      1510       1520       1530       1540       1550       1560       1570       1580       1590       1600
AGATAGATAG ATACATAGAT AGATAGATAG ATAGACAGAC AGACATATGA TAGATAGATA GATAGACAGA CAGACATATG ATAGACAGAT AGATAGACAG

      1610       1620       1630       1640       1650       1660       1670       1680       1690       1700
ACAGACATGA TAGATAGACA GACAGACATG ATAGATAGAT AGATAGATAG ATAGATAGAT AGATAGATAG ATAGATAGAT AGATAGATGA TAGATAGATA

      1710       1720       1730       1740       1750       1760       1770       1780       1790       1800
GATAGATAGA TGATAGACAG ACATATGATA GATAGATAGA TAGATAGATA GATAGATAGA TAGATAGATA GATATAGACA GACAGACATA TGATAGACAG

      1810       1820       1830       1840       1850       1860       1870       1880       1890       1900
ATAGATAGAC AGACAGACAT ATGATAGATA GATAGATAGA TAGATAGATA TAGATAGATG ATAGATAGAC AGACATATGA TAGATAGATA GATAGATAGA

      1910       1920       1930       1940       1950       1960       1970       1980       1990       2000
GAGAGAGAGA GAGAGAAAGA GAGAGATAGA TAGATAGATA GATAGATAtg ttcagtaaac atatgatcca attatttta gtcacaactc taaaattgtt
```

b) GOE(Mouse2)

```
        10         20         30         40         50         60         70         80         90        100
ccgttcggaa agaagatatt tagttttaga agtacagaag atcaatatga gagtttctag agtagatgaa agagacaatc tagactcatc tgtaagtaat

       110        120        130        140        150        160        170        180        190        200
gttcatttgt aaaattcata aacttttagc aattagtaat tctaGATAAA TGATAGATAG ATAGACAACA GAAGATACAT AGATAGATAG ATAGATAGAT

       210        220        230        240        250        260        270        280        290        300
AGATAGATAG ATAGATGATA GATAGATGAT AGATAGATAG ATAaattaac acgtagggag gtagatagac ttagacagag cattcagagt gactatgtaa

       310        320        330        340        350        360        370        380        390        400
tattaatttt tgcatgagta ctgacttgta tctgagagtt tgatgcagta ttatatgtga cacagtatat tgtgatagaa ctacgattag tcaagacagc

       410        420
taaaattact actacgaaga
```

c) GOE(Rat)

```
        10         20         30         40         50         60         70         80         90        100
agctagaaag accattagat ggttggcaaa taGATAGATA GATAGATAGA TAGATAGATA GATAGATAGA TAGATAGATA GATAAAAGCC CTAAGAAGAC

       110        120        130        140        150        160        170        180        190        200
AGTATGTTAA TAGATAACTA GAAAGATCAT TAGATTGTTG GTAGACAGAT AGTTAGGTAG ATAGATAGAT AGATAGATAG ATAGATAGAT AGATAGATAG

       210        220        230        240        250        260
ATAGATAGAT AGATATGTAA TAGATGGATG GATAGATAGA TAGATAGATA gacagacaga c
```

d) GOE(Snake)

```
      1460       1470       1480       1490       1500       1510       1520       1530       1540       1550
tatctggtgc aaatattttt ttaaaaaaga ctccaataat catggtaagc taaatcagct caggatataa gtctggaagc GATAGATAGA TAGATAGATA

      1560       1570       1580       1590       1600       1610       1620       1630       1640       1650
GATAGATAGA TAGATAGATA GATAGATAGA TAGATAGATA GACAGACAGA CTAAAAGCTA TAGATAGATA GATAGATAGA TAGATAGATA GATAGATAGA

      1660       1670       1680       1690       1700       1710       1720       1730       1740       1750
TAGACAGACA GACAGACAGA CAGACAGACA GACAGACAGA CAtttgaaga ggttacttat taatatagaa ggaaatttct gccaatacaa gcaattgttg
```

Figure 5.7

Alignment of the 15 nucleotides on either side of the GATA regions in the four non-*Drosophila* GOE sequences. At only one position is there the same nucleotide in all four sequences (boxed).

Sequence	5' region	GATA region	3' region
GOE(Mousel)	CCCCACACGCAGACA	.....	TGTTCAGTAAACATA
GOE(Mouse2)	CAATTAGTAATTCTA	.....	AATTAACACGTAGGG
GOE(Rat)	GATGGTTGGCAAATA	.....	GACAGACAGAC
GOE(Snake)	TATAAGTCTGGAAGC	.....	GATGAAGAGGTTACT

Chapter 6

ARRANGEMENT OF GOE AND FLANKING SEQUENCES IN *DROSOPHILA* GENOMES.

6.1 Genomic landscapes of GOE6 and GOE5

It is apparent from the sequence data that GOE elements are unlike any of the transposable repeated sequences that were discussed in Chapter 1. Yet GOE elements are dispersed in the genome. This dispersed arrangement can be explained in at least three ways:

i) GOE elements may represent a novel class of mobile sequence. This is unlikely, however, since they are not flanked by short direct repeats. These short direct repeats are thought to arise by the duplication of host sequences upon insertion of a transposable element.

ii) GOE elements themselves may be immobile but are part of a larger species of transposable element.

iii) the apparent dispersal reflects instead a more permanent state of affairs. GOE elements are as immobile as conventional genes. Either they arose *in situ* or they were dispersed by a series of chromosomal rearrangements, such as inversions or translocations.

These last two points can be approached primarily by

looking at the arrangement of unique and repeated sequences (i.e. the genomic landscape) around the GOE elements.

Unique and repeated sequences are not distributed uniformly throughout the euchromatin. For example, extensive chromosomal 'walks' that cover up to 100kb in some regions have encountered few repeated sequences (e.g. Spierer et al., 1983). On the other hand, some regions may be quite rich in repeated sequences (e.g. region 19F as described by Miklos et al., 1984). Repeated sequences are also predominantly located in regions of intercalary heterochromatin (Ananiev et al., 1979 and Zhimulev et al., 1982). Apart from region 19F-20AB, the known cytological locations of GOE elements do not correspond to those locations that are regarded as intercalary heterochromatin by the above authors. To investigate whether GOE elements are associated with other repeated sequences, *D. melanogaster* genomes were probed with plasmid subclones that together cover all the cloned flanking sequences on either side of GOE6 and GOE5. Both Canton S DNA and DNAs from three wild-type strains were probed (the strains are described in the legend to figure 6.1).

Figure 6.1 presents the results obtained when these genomic DNAs were separately probed with six plasmid subclones that cover the sequences around GOE6.

p48-11 and p48-3 both hybridise to single genomic restriction fragments, indicating that they carry unique sequences. Plasmids p41-20, p48-4 and p316-D all hybridise to several restriction fragments in all four genomes and so

contain repeated sequences. Since the patterns of hybridisation for the three plasmids are different, they probably contain non-homologous repeated sequences. This is confirmed by the fact that the insert DNAs from these three plasmids do not cross-hybridise (not shown). Some differences are seen in the hybridisation pattern of p41-20 to the four strains, which would indicate either that their are restriction site polymorphisms or that the repeated sequence is mobile. Two other plasmids, p41-11 and p316-B9, hybridise to single bands and so probably contain unique sequences.

Four subclones flanking GOE5 (p17B, p315-11B, p17C and p315-8) were separately probed to Eco RI digested Canton S DNA only. Except for p17C, they all hybridise to unique restriction fragments. From the pattern of hybridisation of p17C to the genome, it is apparent that it is not related to any of the repeated sequences that flank GOE6.

Since these two GOE elements are situated adjacent to unique regions they cannot be parts of a larger repeated sequence. Furthermore, the arrangement and type of repeated sequences that flank these GOE elements are not the same. This indicates that they are apparently situated in unrelated regions of the genome. It is most likely that GOE elements have not been transposed to their present locations. However, if GOE elements and their surrounding sequences are flanked by a pair of transposable elements, the entire unit could be mobilised. For example, a large transposable element, called TE1, contains the *white* and *roughest* genes and is thought to

be terminated by *copia* sequences, which are responsible for its mobility (Ising and Block, 1980).

6.2 The hybridisation pattern of GOE elements to different *Drosophila* genomes.

Although sequences homologous to Bkm DNA (and therefore presumed to be equivalent to GOE) are present in a wide range of organisms (e.g. Jones and Singh, 1982), it is not known how well the numbers of GOE elements are maintained in a set of related species. Middle repeated sequences in *Drosophila* are usually present in only a few closely related species (Dowsett and Young, 1982 and Dowsett, 1983). A selection of *Drosophila* genomes were therefore probed with GOE elements to test their between species distribution. The probe used was a 450bp Hae III restriction fragment from p315-T22, which was first cloned into the Hinc II site of single-stranded phage, M13mp8 (Messing and Vieira, 1982). This was sequenced and shown to contain a continuous run of 17 GATA tetranucleotide units (see Chapter 4). The *Drosophila* restriction fragment was liberated by digesting the double stranded form of the M13 recombinant with Eco RI and Hind III (which span the Hinc II site), and then re-cloned into the corresponding sites of pBR322. This plasmid recombinant was designated pGOE5. Radioactively-labelled pGOE5 probes were prepared using alpha-^{32}P-dATP, as opposed to alpha-^{32}P-dCTP, as this increases the potential for

incorporation of radioactive nucleotides into the GATA region three-fold.

6.2.1. Hybridisation of pGOE5 to male and female *Drosophila melanogaster* (Canton S) DNA.

As GOE has been implicated in sex differentiation, the hybridisation of GOE to male and female *Drosophila* was compared to see if any sex differences are apparent. Aliquants of DNA from the heads of male and female flies were digested to completion with one of the following enzymes: Hinf I, Sau 3AI or Alu I. After separation in agarose gels and Southern-blotting to nitrocellulose filters, the DNA was probed with pGOE5. The results are shown in figure 6.2.

The positions of the discrete hybridising restriction fragments are identical in both sexes, and intense hybridisation is seen to high molecular weight DNA for all enzymes used (table 6.1 overleaf lists the sizes of the restriction fragments that hybridise to pGOE5 in all the genomes tested). The intensity of the signal to the high molecular weight fraction is less in males than in females. (The intensity of hybridisation to the discrete restriction fragments is too light to make a similar comparison). The autoradiograph film is saturated in its response, and it would be inappropriate to compare densities quantitatively. That less hybridisation is seen in the male indicates that some, at least, of the high molecular weight fraction is resident on the X chromosome.

Table 6.1. Sizes of GOE restriction fragments in various *Drosophila* genomes

Genome	Enzyme	Restriction fragment sizes (kb)
CS	Hinf I	2.25, 1.45, 1.25.
CS	Alu I	2.00, 1.20, 1.10, 0.62, 0.58, 0.36
FF	Alu I	1.60, 1.30, 1.05, 0.76, 0.70, 0.48
D. virilis	Alu I	1.00, 0.94, 0.82, 0.72, 0.58, 0.50, 0.41
D. pseudo.	Alu I	2.15, 1.70, 1.55, 1.45, 1.30, 0.98, 0.68

As the pattern of hybridisation is similar in both sexes one can conclude that GOEs are not involved in any hypothetical structural rearrangements that may accompany sex differentiation. This idea was initially proposed by Singh et al., (1980b) to account for the different hybridisation patterns of Bkm to male and female mice.

6.2.2. Hybridisation of pGOE5 to the Canton S and FF strains *of Drosophila melanogaster*.

Differences in hybridisation pattern of Bkm to the genomic DNAs of some *D. melanogaster* strains have been noted by Jones and Singh (1981). This was further investigated by comparing Canton S DNA with DNA from a wild-type strain. Strain FF is one of a number of lines (set up from single matings) that are homozygous for the second chromosome and contain the Adh^F allele (Lewis and Gibson, 1978 and see section 6.2). Though the other chromosome pairs are likewise derived from the same wild population, they may, or may not, be homozygous.

Embryonic DNA from FF, and adult DNA from Canton S flies were digested to completion with either Eco RI, Pst I, Hinf I, Sau 3AI or Alu I enzymes. It has already been shown by Singh et al., (1980b) that whole adult and embryonic DNAs give identical patterns of hybridisation when probed with Bkm DNA. Therefore, if these two DNAs show different patterns of hybridisation, they can be attributed to differences between the genomes of the two strains. Southern-blotted restriction

fragments were probed with pGOE5 and these results are presented in figure 6.3.

Almost all positions of the discretely hybridising restriction fragments are different between Canton S and FF, for all five enzymes, though both strains possess the high molecular weight component in roughly equal amounts. The most striking difference is the absence from FF of a relatively heavily labelled 2.0kb Alu I restriction fragment that is present in Canton S. The intensity of the signal in Canton S may be due to the presence of several 2.0kb Alu restriction fragments containing a single copy of GOE, to the presence of a single Alu restriction fragment containing several copies of GOE, or a combination of the two.

That the 2.0kb restriction fragment is absent from the FF strain suggests either that the GOE elements themselves are absent, or restriction enzyme site changes have generated a larger GOE-containing Alu restriction fragment that is now hidden amongst the intense labelling associated with hybridisation to the high molecular weight restriction fragments.

In summary, although the sizes of the discrete GOE restriction fragments are different between Canton S and FF, there are similar amounts both of these and of the high molecular weight component in the two strains.

6.2.3 Comparison of the hybridisation patterns of GOE elements from the Canton S and FF strains

The two FF GOE copies (in plasmids pFF3 and pFF12-2) were isolated on the basis of hybridisation to GOE6, which was chosen because of its long, unbroken stretch of GATAs. A comparison was made between the hybridisation pattern of the plasmids, pGOE5 and pFF12-2 to the Canton S and FF genomes and to the other GOE elements that had been cloned and isolated. All the DNAs were digested with Eco RI and Alu I enzymes. The plasmid subclones are genomic Eco RI restriction fragments and the location of the GOE elements relative to the Eco RI and Alu I sites were not known at this stage. This ensures however that the plasmid restriction fragments hybridising to GOE sequences will migrate to the same position in the gel as their genomic equivalents.

Figure 6.4 presents the results. Although the discrete, lower molecular weight restriction fragments are only lightly labelled, it can be seen that the two GOE elements show similar hybridisation patterns, indicating that the same genomic restriction fragments are hybridising to both GOE elements.

This experiment has relevance to the previous discussion on the number of GOE copies in *Drosophila melanogaster*, and whether all had been isolated (section 3.4). The positions of the GOE-containing restriction fragments from GOE4, 6, 8 and 9 correspond well with four of the restriction fragments in the Canton S genome. The GOE5 and 7 bands do not correspond to

any visible restriction fragments, but these may be too lightly labelled to be detected. Only one visible genomic restriction fragment does not correspond to the cloned GOEs and that is the relatively heavily labelled 2.0kb Alu restriction fragment. Assuming that the visible genomic restriction fragments (as well as the two undetectable ones that probably correspond to GOE5 and 7) are the only ones that comprise the discrete or low molecular weight component of GOE, then 6 out of the 7 dispersed genomic copies have been isolated as lambda and plasmid clones.

6.2.4. Hybridisation of pGOE5 to the genomes of different *Drosophila* species.

Genomic DNAs from four species, representing three of the six subgenera of the genus Drosophilidae (Patterson and Stone, 1952), were analysed with respect to hybridisation to pGOE5. DNAs from *D. melanogaster* (subgenus: Sophophora), *D. pseudoobscura* (subgenus: Sophophora), *D. virilis* (subgenus: Drosophila) and *D. silvarentis* (suborder: Hirtodrosophila) were digested to completion with excess Alu I enzyme, separated on agarose gels and Southern-blotted. Filters were probed with pGOE5 and results are shown in figures 6.5 and 6.6. The most striking differences between *D. melanogaster* and the others is the complete absence of hybridisation to high molecular weight restriction fragments in the latter species. The DNAs however are not all from diploid tissue and if the high molecular weight GOE elements reside in

heterochromatin then it can be argued that the relative underreplication of heterochromatic sequences in polytene tissue could account for this apparent absence of signal. However, both whole adult (predominantly polyploid) and embryonic (diploid) *D. melanogaster* DNA give identical patterns of hybridisation with GOE elements (Singh et al., 1981b using Bkm DNA). Also, DNA from both *D. virilis* adults and ovaries (which have more than a diploid amount of satellite sequences at eclosion (Endow and Gall, 1975)) show identical patterns, indicating that the bulk of *D. virilis* centromeric heterochromatin and satellite DNAs do not contain GOE elements. The conclusion is that high molecular weight GOE elements are absent from the genome of *D. virilis*, and probably from that of *D. pseudoobscura* as well.

With the exception of *D. silvarentis*, all the genomes possess visible and discrete lower molecular weight GOE restriction fragments. GOE may still be present in this genome but less abundantly.

6.3 Summary of hybridisation experiments with pGOE5.

In summary, species from four subgenera of Drosophilidae possess discrete low molecular weight GOE restriction fragments in similar amounts (6-8 hybridising bands are detectable in each genome). This is in contrast to the situation for most middle repeated sequences, which tend to be

limited to a few species within a subgenus. This also shows that the widespread distribution of GOE between phyla can also be reflected within a genus. As high molecular weight GOE is present only in *D. melanogaster*, whereas the discrete restriction fragments are present throughout the genus, the two *D. melanogaster* GOE components can be treated separately. Even though the high molecular weight component represents the bulk of GOE-hybridising sequences in this genome, it is the low molecular weight component that is most likely to have a function in the genome, because this is maintained in other *Drosophila* species. It is this component that has been sequenced and analysed here (Chapters 4 and 5). The possible functions of the GOE family of repeated sequences must therefore be considered in the light of the evidence that GOE elements apparently arose by the accumulation of random mutation events in a poly(GATA) sequence. The effect that this has on the proposition that GOE elements are involved in sex determining processes (section 1.4) is discussed in the final chapter.

Figure 6.1

Hybridisation of DNA sequences flanking GOE6 to *D. melanogaster* genomes. 1-2ug of embryonic genomic DNAs from the following *D. melanogaster* strains were digested to completion with 5 units of Eco RI enzyme and separated on 1% agarose gels for 16 hours: Canton S (lane CS), Fr (lane A), FF (lane B) and SS (lane C)*. Three Southern blots were prepared from three sets of digestions. Filter were probed with radioactively-labelled DNAs as indicated in the figure. Filters were also probed with radioactively-labelled lambda C1857 DNA in order to reveal the marker fragments. Autoradiograms were exposed for 3-7 days. The restriction map shown combines parts of the maps of lambda clones 316, 48 and 41 (see figure 3.6).

*FF and SS are separate lines derived from individuals collected in New South Wales and carry the fast (Adh^F) and slow (Adh^S) alleles of the *alcohol dehydrogenase* locus, respectively (Lewis and Gibson, 1978). Fr is a strain derived from individuals collected in Iowa, USA and which carries a heat resistant allele of the same locus (Sampsell, 1977). These three strains are homozygous for the second chromosome, which carries the *Adh* locus. The other chromosomes of the FF and SS strains are also derived from the same wild population, but may or may not be homozygous. Homozygosity for the second chromosome of the Fr strain was achieved by crosses with balancer stocks and so homozygosity for the other chromosomes cannot be assumed.

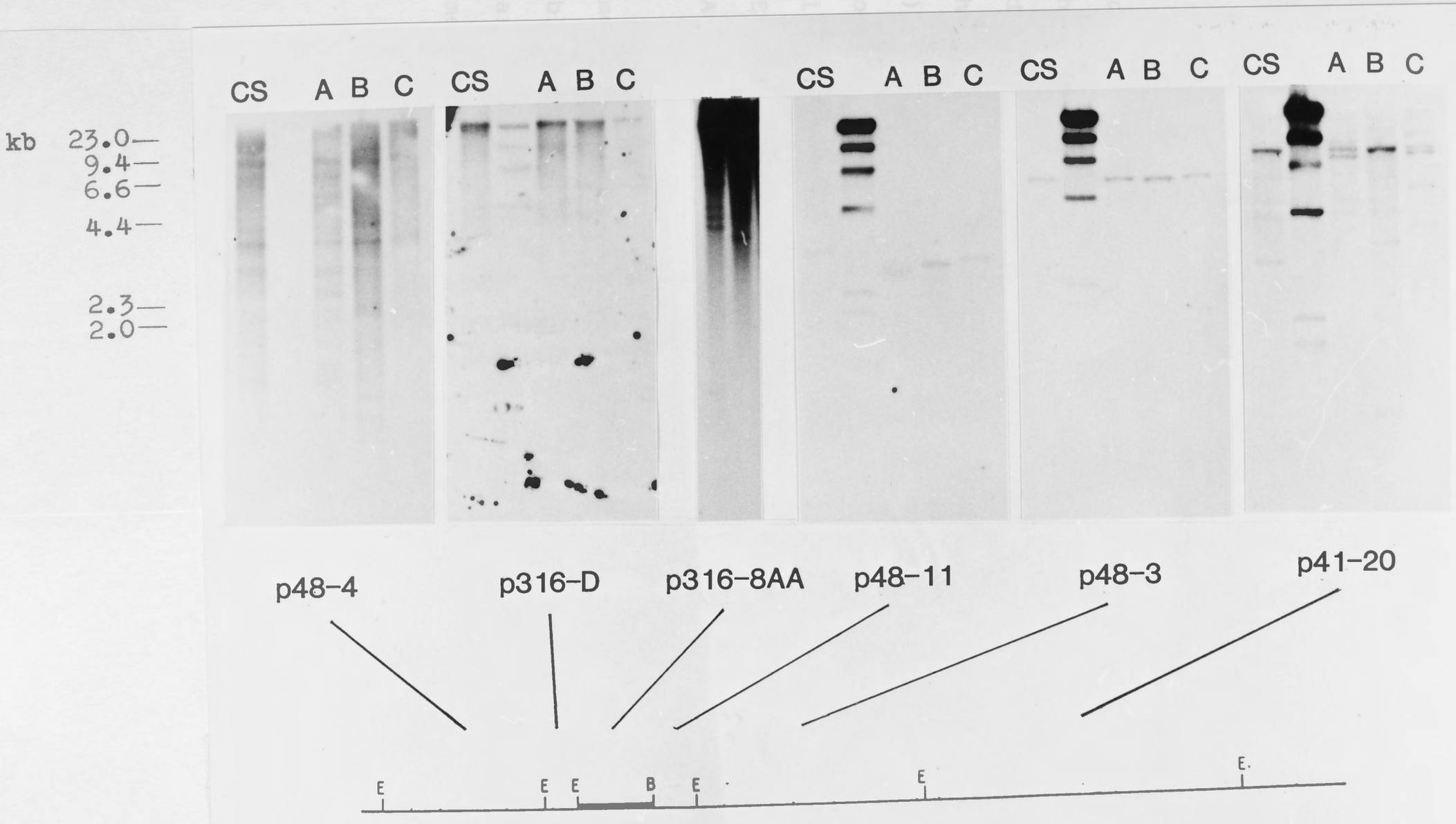

CS A B C CS A B C CS A B C CS A B C CS A B C
kb
23.0
9.4
6.6
4.4
2.3
2.0
p48-4
p316-D
p316-8AA
p48-11
p48-3
p41-20
E
E E
B
E
E
E

Figure 6.2

Hybridisation of pGOE5 DNA to male and female adult *Drosophila melanogaster* genomic DNAs. 2ug each of DNA from the heads of male (lanes 1,3 and 5) and female (lanes 2,4 and 6) adult flies were digested to completion with 5 units of one of the following enzymes; Hinf I (Hinf on the figure), Sau 3AI (Sau) or Alu I (Alu). Fragments were separated on a 1% agarose gel, Southern blotted and probed with radioactively-labelled pGOE5 DNA.

a) Ethidium bromide staining pattern under UV light.

b) Autoradiogram of filter after 2 weeks' exposure. Apart from the strong hybridisation to the high molecular weight fragments, discrete lightly labelled fragments are also visible in all lanes, lying between 2.5kb and 0.5kb in size. The arrow points to the relatively heavily-labelled 2.0kb Alu fragment that is mentioned in the text.

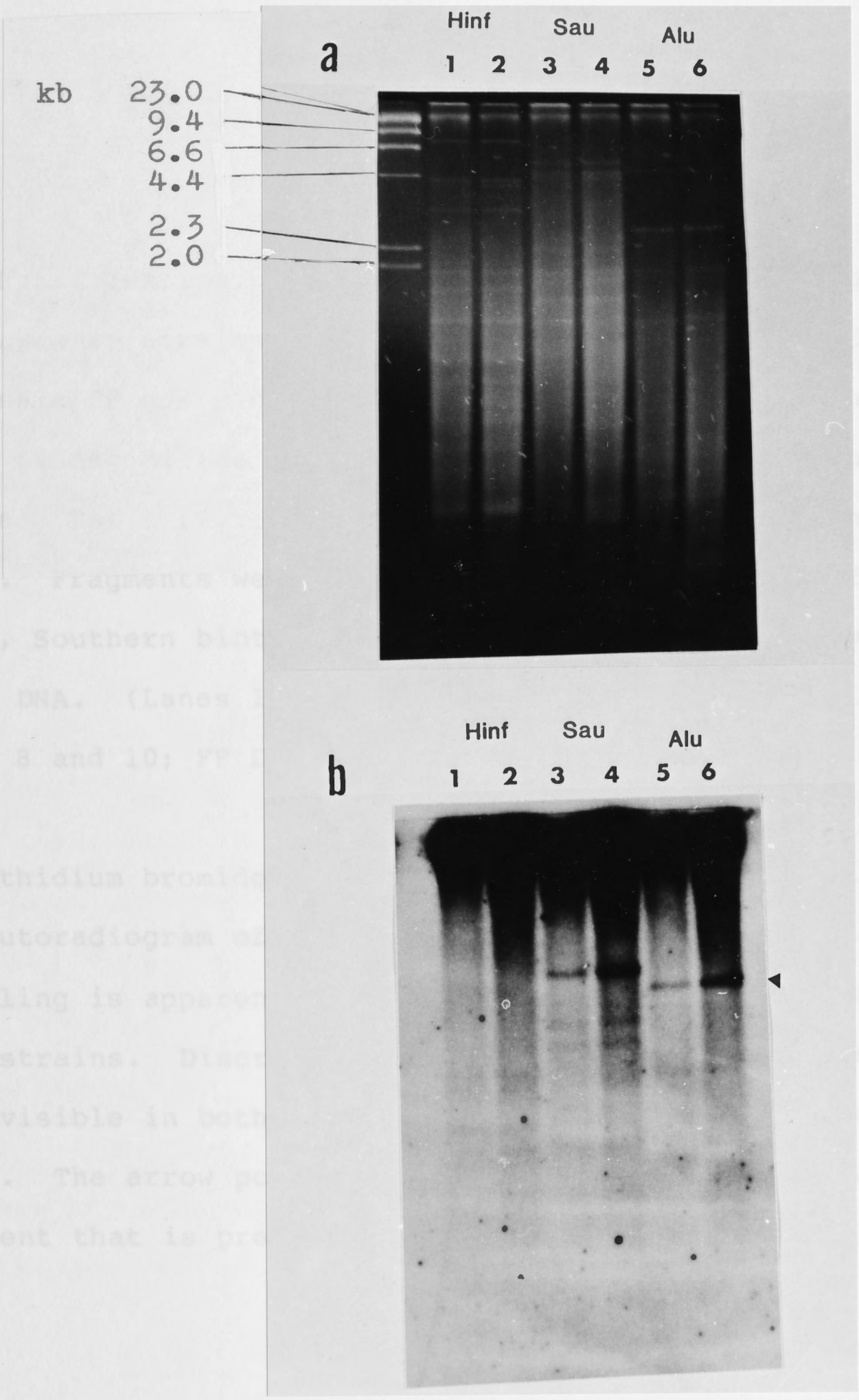

a
Hinf
Sau
Alu
1 2 3 4 5 6
kb
23.0
9.4
6.6
4.4
2.3
2.0
b
Hinf
Sau
Alu
1 2 3 4 5 6

Figure 6.3

Hybridisation of pGOE5 to genomic DNA from two *Drosophila melanogaster* strains. 2ug each of adult Canton S and embryonic FF genomic DNAs were digested to completion with 5 units of one of the following enzymes; Eco RI (Eco in the figure), Pst I (Pst), Hinf I (Hinf), Sau 3AI (Sau) and Alu I (Alu). Fragments were separated on 1% agarose gels for 16 hours, Southern blotted and probed with radioactively-labelled pGOE5 DNA. (Lanes 1, 3, 5, 7 and 9; Canton S DNA. Lanes 2, 4, 6, 8 and 10; FF DNA).

a) Ethidium bromide staining pattern under UV light.

b) Autoradiogram of filter after 1½ weeks' exposure. Intense labelling is apparent to high molecular weight fragments in both strains. Discrete lower molecular weight fragments are also visible in both strains, but they are of different sizes. The arrow points to the intensely labelled 2.0kb Alu I fragment that is present in Canton S but absent in FF.

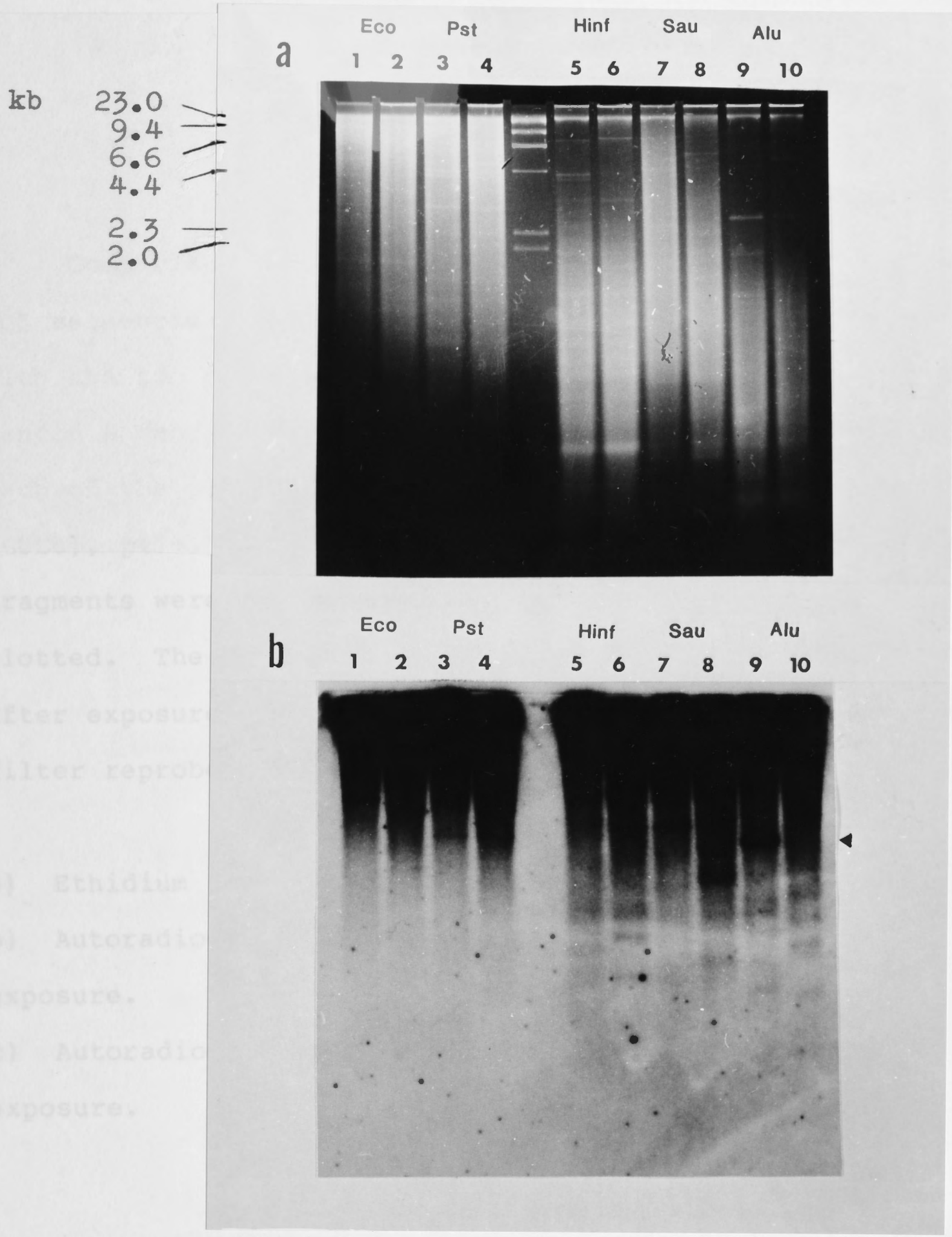

kb
23.0
9.4
6.6
4.4
2.3
2.0
a
Eco
Pst
Hinf
Sau
Alu
1 2 3 4 5 6 7 8 9 10
b
Eco
Pst
Hinf
Sau
Alu
1 2 3 4 5 6 7 8 9 10

Figure 6.4

Comparison of hybridisation pattern of Canton S and FF GOE sequences. The following DNAs were digested to completion with the two enzymes, Alu I and Eco RI, in combination; 3ug Canton S embryonic DNA (CS), 3ug FF embryonic DNA (FF), 0.1ug each of the plasmids, p1-2 (GOE4), p315-11 (GOE5), p48-13 (GOE6), p47-13 (GOE7), p28A (GOE8) and p319-13(GOE9). The fragments were separated on a 1% agarose gel and Southern blotted. The filter was probed first with pGOE5 DNA, which after exposure to autoradiographic film was washed off and the filter reprobed with pFF12-2 DNA.

a) Ethidium bromide staining pattern under UV light.

b) Autoradiogram of pFF12-2 hybridisation after 3 weeks' exposure.

c) Autoradiogram of pGOE5 hybridisation after 4 weeks' exposure.

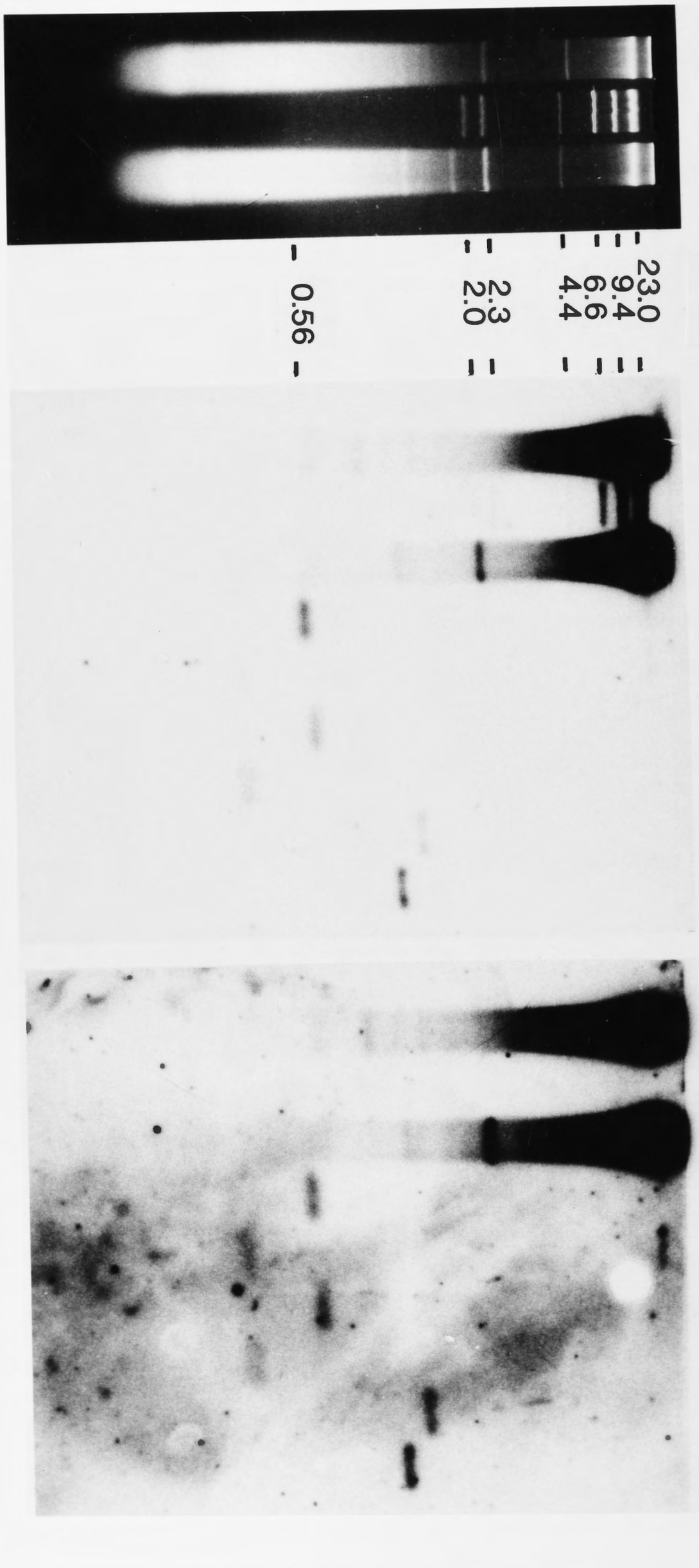

FF
CS
23.0
9.4
6.6
4.4
2.3
2.0
0.56
KB
FF
CS
GOE4
GOE5
GOE6
GOE7
GOE8
GOE9
FF
CS
GOE4
GOE5
GOE6
GOE7
GOE8
GOE9

Figure 6.5

Hybridisation of pGOE5 DNA to *D. melanogaster* and *D. virilis* genomic DNAs. 5ug of genomic DNA isolated from *D. virilis* (vir) ovaries and 5ug and 0.5ug of genomic DNA isolated from *D. melanogaster* (mel) adults were digested to completion with 5 units of Alu I restriction enzyme. Fragments were separated on a 1% agarose gel, Southern blotted and probed with radioactively-labelled pGOE5 DNA. Lambda C1857 DNA, digested with Hind III, provided size markers.

a) Ethidium bromide staining pattern under UV light.

b) Autoradiogram of filter after 2 weeks' exposure. No heavy labelling to high molecular weight fragments is apparent in *D. virilis* DNA, even when one tenth of the amount of *D. melanogaster* DNA shows it clearly. *D. virilis* does possess 6-8 discrete lower molecular weight fragments, ranging in size from 2.0 to 0.5kb.

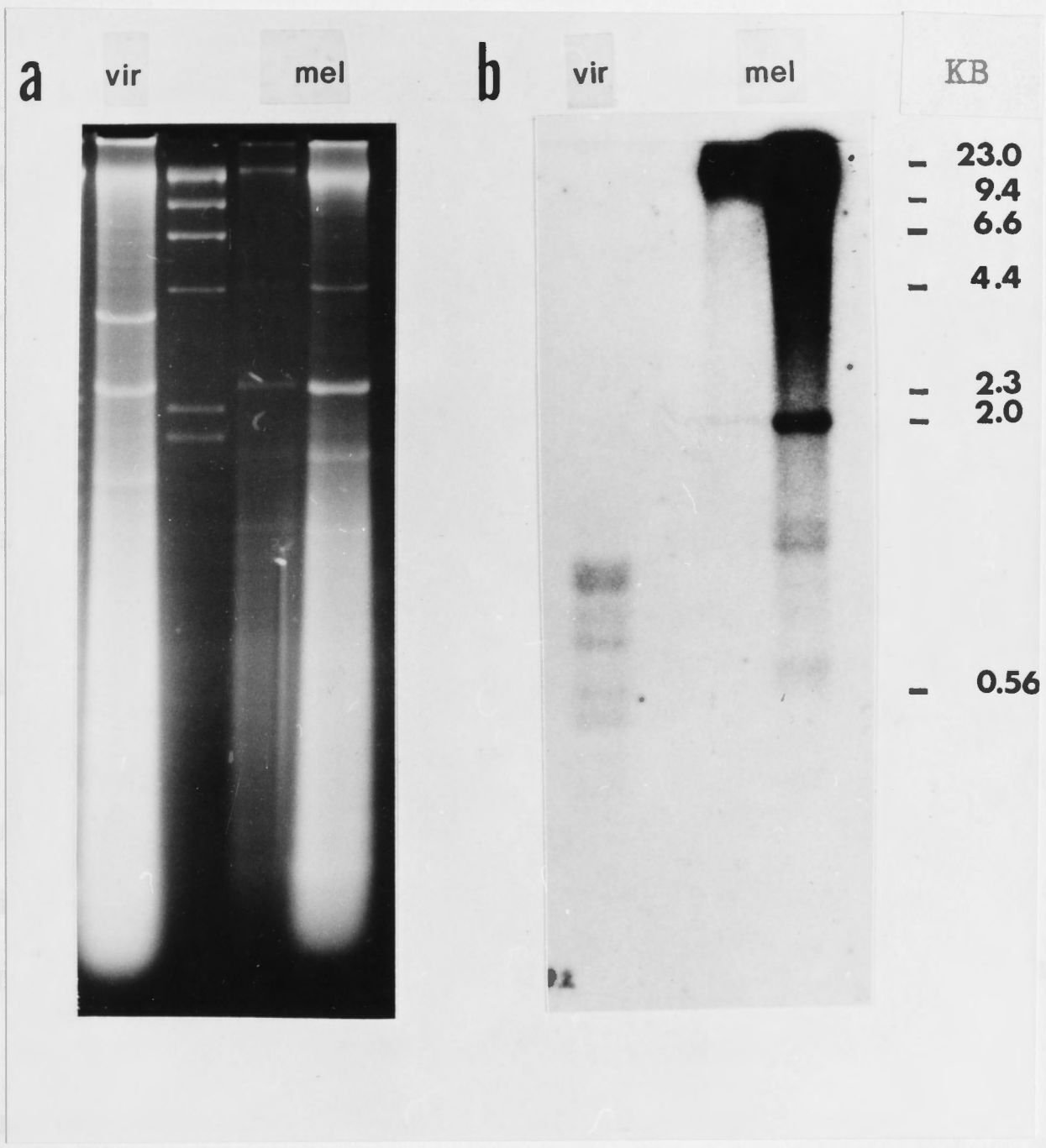

a
vir
mel
b
vir
mel
KB
23.0
9.4
6.6
4.4
2.3
2.0
0.56

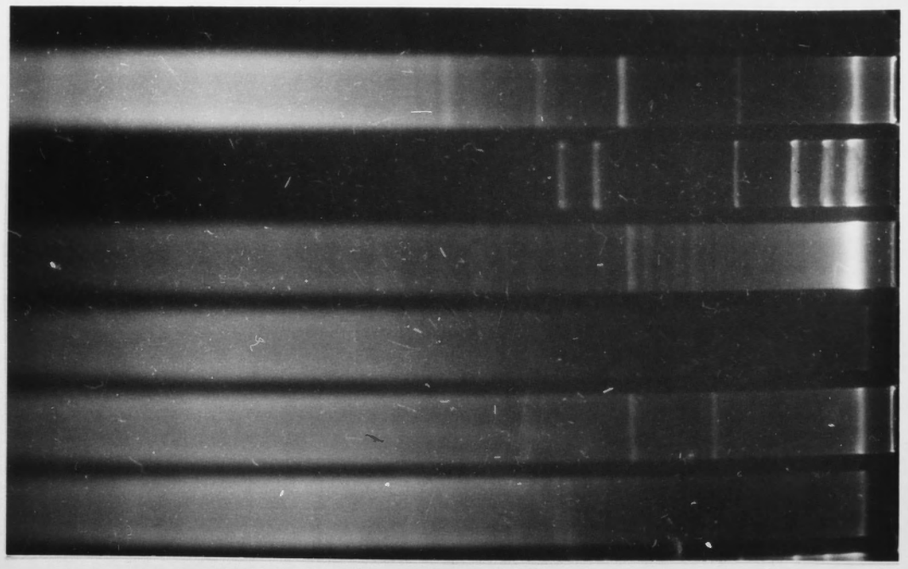
Figure 6.6

Hybridisation of pGOE5 DNA to genomic DNAs from four *Drosophila* species. 2-3ug of the following genomic DNAs were digested to completion with 5 units of Alu I enzyme; *D. melanogaster* adult heads (mel in the figure), *D. silvarentis* adult heads (silv(h)), *D. silvarentis* salivary glands (silv(s)), *D. virilis* ovaries (vir) and *D. pseudoobscura* whole adult bodies (pseud). Fragments were separated on a 1% agarose gel for 16 hours, Southern blotted and probed with radioactively-labelled pGOE5 DNA.

a) Ethidium bromide staining pattern under UV light.

b) Autoradiogram of filter after 2 weeks' exposure. Note the absence of hybridisation to high molecular weight fragments in all species except *D. melanogaster*. Low molecular weight discrete fragments are visible in all species except *D. silvarentis*.

a.



mel

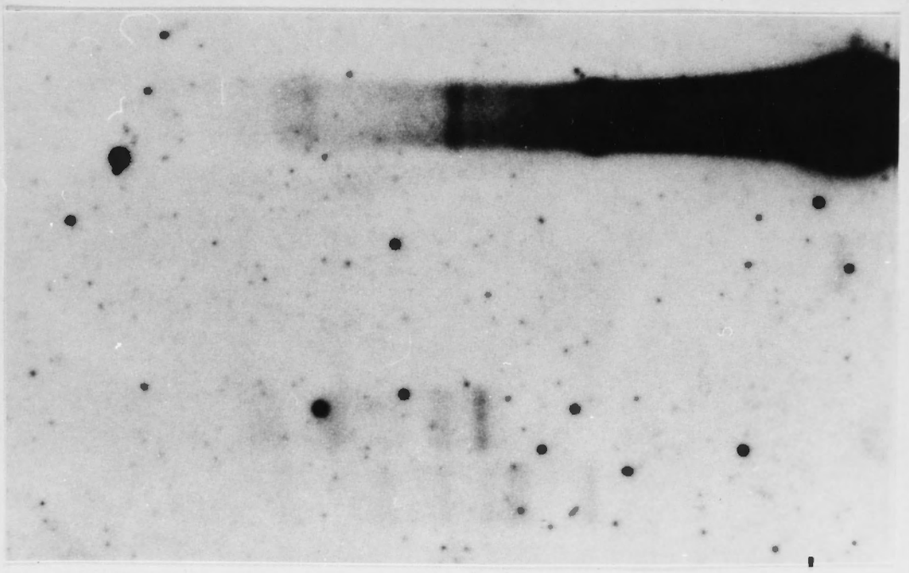
silv(h)

silv(s)

vir

pseud

b.



mel

silv(h)

silv(s)

vir

pseud

KB

23.0
9.4
6.6
4.4
2.3
2.0
0.56

Chapter 7

DISCUSSION

The genomes of three *Drosophila* species (*D. melanogaster*, *D. pseudoobscura* and *D. virilis*) each contain eight to ten dispersed copies of the Garden of Eden (GOE) family of middle repeated sequences. Five different copies from the *D. melanogaster* genome have been sequenced and analysed here (Chapter 5). Structural data for a major proportion of a family of repeated sequences has therefore been obtained - an essential step for elucidating function. Furthermore, this information can be used to test the validity of the proposals (Jones, 1983, Jones and Singh, 1981, Singh et al., 1980b, 1984 and Epplen et al., 1982, 1983a,b), that GOE elements are involved in major developmental events, especially sex determination.

7.1 GOE elements and sex determination

Three sets of evidence have been presented to support the claim that GOE elements are somehow involved in the processes of sex determination (see Introduction, section 1.4):

a) GOE elements are 'conserved' over a wide range of eukaryote species,

b) GOE elements hybridise *in situ* more intensely to the sex chromosomes than to the autosomes of snakes, birds and mice.
c) GOE elements, or their adjacent sequences, hybridise specifically to some mouse male-specific RNA molecules.

Each of these points can now be examined in more detail, with reference to the structural data that is available for *Drosophila*, snake and mouse GOE elements.

7.1.1 The 'conserved' nature of GOE elements.

The apparent conservation of nucleotide sequences suggests that selection has acted against the accumulation of mutations that would otherwise disrupt an essential function. Mechanisms exist that can homogenise a particular repeated sequence family in the apparent absence of selective constraints. This process of 'molecular drive' (Dover, 1982) serves to explain why different species can have different families of repeated sequences, which are however quite homogeneous. GOE elements, however, do not show this characteristic. For example, more variation is seen between copies within a population than between populations (compare, for example, the sequence of GOE6 to those GOE4, 5 and 9 and to the sequence of GOE6 from the FF strain). Therefore, molecular drive is unlikely to be able to explain the apparent conservation of GOE elements in eukaryote genomes. Does the fact that GOE elements hybridise to many different genomes then mean that their sequences have been maintained in order to preserve an encoded function? In other words, are GOE

elements really conserved?

Suppose that a family of repeated sequences encodes some essential function. Either the entire sequence, or specific parts of it, must be maintained. For example, a family of structural genes should preserve the nucleotides at the first and second codon positions better than those at the third codon positions. Assuming that the original GOE element is most likely to be a tandem array of GATA tetranucleotides, the analyses of Chapter 5 have shown that the *Drosophila* GOE elements could have arisen by the accumulation of random nucleotide changes within poly(GATA) sequences. Furthermore, the lengths of the GOE elements are all different. Thus, within the *Drosophila* genome, the only real similarity between GOE elements is a localised distribution of GATAs. The conserved aspect of GOE elements, between species, is very unlikely to be more than that within a species. For example, the two GOE elements from the mouse, and the GOE element from the snake are of different lengths and have different proportions and distributions of GATA units (see figure 5.1). Again, a localised distribution of GATA units is the only common characteristic between these GOE elements.

What has been described as a conserved sequence in many eukaryote genomes is really clusters of GATA tetranucleotides. How many contiguous GATAs would be needed to be detected as GOE element? This will depend on the hybridisation conditions and on the number of GATA units in the probe. The smallest GOE element (GOE5) detected here

contains 21 GATAs, 17 of which are contiguous. This was used to detect similar sequences in other *Drosophila* genomes. Bkm is a satellite DNA and so probably contains many more contiguous GATA units than the GOE5 sequence. More of it is therefore potentially capable of hybridising with poly(GATA) sequences and should give a stronger hybridisation signal. Also, the hybridisation conditions used by Singh and Jones (1981) (60-62°C) are less stringent than those used here. Therefore, some of the similar sequences that they detected in other genomes probably had fewer than 17 contiguous GATA units. Stable duplexes of length 30-50 nucleotides can form under these conditions (McCarthy and Church, 1970) In other words, a stretch of as few as 7-12 contiguous GATAs could be classed as a GOE element.

A class of sequences, whose only common characteristic is that each possess at least 7-12 contiguous GATA units, cannot strictly be considered as being structurally conserved. The conclusion from this section is that the apparent conservation of GOE elements in many eukaryote genomes is not a sufficient criterion for assuming that they have some fundamental role to play.

7.1.2 Association of GOE elements with sex chromosomes.

The Y sex chromosome of mammals and the W sex chromosome of birds are essential for differential sex development and must contain either the structural or the regulatory locus for a sex determining gene. The predominant site of hybridisation

of GOE elements (Bkm) is to these two chromosomes. This suggested to Singh and Jones that GOE elements could either be the sex determining genes themselves or that they are in some other way involved in the sex determination process.

An initial suggestion was that GOE elements be involved in the gradual heterochromatinisation of one of the sex homologue chromosomes. This would lead to the creation of a W (or Y) chromosome (Singh et al., 1980a). There are other satellite sequences that are localised solely on sex chromosomes. For example, there are the 3.4 and 2.1kb satellite sequences found predominantly on the human Y chromosome (Cooke et al., 1976) and the W-chromosome specific satellite sequences of the chicken (Tone et al., 1982 and 1984). Unlike GOE elements, these satellite sequences are not readily detectable in related species. GOE elements are not the sole example of sex-chromosome specific sequences, though the fact that they are generally found on the sex chromosomes of several species might suggest that they are more important.

The molecular details underlying heterochromatinisation are not known. Several sequences might be amplified in the process, without being directly involved in the process itself. Since GOE elements are dispersed on most chromosomes they were probably also located on a proto-W or proto-Y chromosome as well. Whenever such a proto-chromosome undergoes heterochromatinisation, GOE elements will be in a position to be amplified along with other sequences. This could explain why GOE elements are relatively concentrated on

sex chromosomes, without supposing that they are directly involved in sex determination.

There are other dispersed and simple sequences that are present in several eukaryotes and probably comprise a class of sequences that would include the GOE elements (see section 7.4). If the above argument is correct, one might expect some of these also to be 'sex chromosome specific'. As far as is known, this has not been tested, but it is certainly feasible to perform the necessary experiment.

Even if GOE elements are involved in the heterochromatinisation process, one would need to explain why the autosomal copies of the GOE element were not amplified along with the sex chromosomal copies. Furthermore, heterochromatinisation is not equivalent to sex determination.

7.1.3 GOE elements and sex-specific transcripts.

The data concerning the transcription of GOE elements are summarised in figure 7.1.

Two distinct functions for the GOE element have been suggested (Epplen et al., 1982, 1983b). The sequence of a snake W-chromosme and GOE-containing clone (pErs5) showed that a GOE element would lie in the 3' non-translated region of a *putative* transcript (Epplen et al., 1982). It could possibly be acting either as a 'signal' or a 'control' sequence. In contrast, it was suggested in a later paper (Epplen et al., 1983b) that the GOE element in a mouse cDNA clone (pmc14) would be *translated*. Singh et al. (1984) proposed that the

GOE element in a mouse genomic clone would also be translated.

In the first case then, it was postulated that the GATA strand of the GOE element was the functional unit (as part of a transcript) while in the second, it was the TATC strand (as the major part of a translated region) that was supposed to be functional.

The results of the transcription studies in the mouse are conflicting in several ways. The snake GOE element hybridised to at least two distinct transcripts in total cellular RNA of liver from male and female mice (one of which corresponds to the mouse cDNA clone, pmc14). However, Singh et al. (1984) report that a *Drosophila* GOE element (GOE4) hybridises only to male liver polysomal RNA. One could resolve these two contrasting results by supposing that GOE elements are indeed transcribed in males and females, but only in the male are these transcripts modified into poly(A)$^+$ RNA. However, GOE elements hybridise to a different population of RNA molecules in male and female brain polysomal RNAs. Therefore, any sex-differentiation in the transcription of sequences that is associated with GOE elements would have had to occur after the differentiation of brain and liver tissue. On this evidence GOE elements need not be involved in sex determination, which is a much earlier event.

Possibly GOE elements are useful for isolating contiguous sequences that are transcribed and/or translated in a sex-specific manner by virtue of the fact that GOE elements are predominantly (though not solely) located on sex

chromosomes. For example, the unique portion of the snake pErs5 clone hybridised to a male-specific mouse transcript (Epplen et al., 1982). However, this would not mean that the GOE elements themselves are primarily involved in sex determination.

The points that were discussed above do not appear sufficient to prove that the GOE elements are important elements involved in sex determination. Of course, neither does this prove that they are not involved. What is known of the processes of sex determination should reveal what properties one can expect from a sequence, or a family of sequences, that is responsible for this major developmental event, and whether GOE elements could fulfil these requirements.

7.2 Sex determination in mammals

Much of what is known of sex determination is derived from studies of mammals, though the essential elements are paralleled in other vertebrates (see McCarrey and Abbott, 1979, for a review). The primary sex determiner is a humoral factor called H-Y antigen. H-Y antigen is thought to be produced in the heterogametic sex of a variety of organisms (e.g. Wachtel et al., 1975). Only gonadal tissue (whether genotypically XX or XY) possesses receptors for H-Y

antigen. In the presence of the antigen the tissue develops into a testis. This organ later produces the hormones that induce development of the male phenotype. In the absence of H-Y antigen, gonadal tissue develops into ovaries, which later direct development of the female phenotype. Consequently, even XX gonadal tissue can develop into testis and so produce a male XX individual, provided H-Y antigen is present early in development. Normally, only the heterogametic sex produces H-Y antigen so that the two sexual phenotypes correspond to the sexual genotypes.*

Either the structural or the regulatory locus for H-Y antigen must reside on the Y chromosome. Loci for other sex-related genes, such as the structural genes for H-Y antigen receptor or for testosterone, reside on the other chromosomes.

If GOE elements are to be important factors in sex determination, could they represent the structural or regulatory loci for H-Y antigen? As has been commented on earlier, the translated product of a GOE (poly(GATA)) sequence would be very hydropobic in nature. It is unlikely to be a soluble, and therefore a humoral, molecule. Possibly a GOE element could be covalently linked to one that can produce a soluble molecule. In which case, the active element is not the GOE element itself.

A GOE element could serve as a regulatory gene, acting either in *cis* or in *trans* with the H-Y antigen structural gene. If it acts in *cis* one would expect repeated copies of the GOE element only if the H-Y antigen structural gene was

* but see addendum
between pp. 202 & 203.

also repeated. If it acts in *trans*, one only needs one copy, irrespective of whether the structural gene is repeated or not. Furthermore, GOE elements are dispersed on all chromosomes (Singh and Jones, 1982), so that the Y chromosomal functional specificity can only be obtained if the non-Y chromosome copies are non-functional. This would lead to a great deal of structural and functional redundancy for a gene that controls one of the earliest and most major switching events in the development of an organism.

7.3 Sex determination in *Drosophila*

Though H-Y antigen is the primary sex determiner in mammals, there could be other areas where GOE elements can be involved in sex determination. Sex determination in *Drosophila* is ostensibly different from that in mammals or birds since it is not dependent on the presence of a Y chromosome. Also, the sex determining factors are not humoral, as they are in mammals, since flies composed of male and female tissue (gynandromorphs) can arise.

Sex determination in *Drosophila* is principally controlled by the ratio of autosomes to sex chromosomes. If the haploid autosomal complement is considered as a unit (A), a chromosomal ratio of 2A:2X produces a female and one of 2A:1X produces a male, whether a Y chromosome is present or not. No specific male-determining or female-determining loci have been

detected. Instead the autosomes and the X chromosome appear to have general 'maleness' and general 'femaleness' specifying functions, respectively. However, there are a number of loci at which the expression of mutant alleles depends on the sex of the fly. These loci are sex-specific, but not necessarily sex determining.

Sex lethal (Sxl) mutants are lethal as homozygotes in females but have no effect as hemizygotes in males. Other mutants at a very closely linked locus are lethal to the male but not the female. It is thought that the *Sxl* product is required for the dosage compensation mechanism that is necessary to ensure that females (XX) and males (X) both produce equivalent amounts of X-linked gene products, possibly by inhibiting expression of the loci on one of the X chromosomes of the female. Constitutive expression of the *Sxl* locus will therefore shut off the only X-linked loci in the male, causing lethality. The expression of *Sxl* is in turn thought to be under the control of the product of the *daughterless, (da)* locus. Females that are homozygous for a *da* mutant fail to produce daughters because the functional product is lacking in the oocyte. Its action is not required by the male. That the *da* product acts in conjunction with the X:A ratio is shown by the fact that triploid intersex progeny survive whereas triploid females die. That is, the measure of the X:A ratio is still the primary determiner of sex. Other sex related loci probably act at later stages. For example, *transformer (tra)* mutants cause 2X:2A individuals to develop a

male phenotype. *Doublesex (dsx)* mutants transform both 1X:2A and 2X:2A flies into identical intersexes, suggesting that it functions to prevent the action of male and female specific development (Baker and Belote, 1983).

The approximate cytological locations of these and other sex loci are listed in table 7.1 (overleaf). It can be seen that none of these correspond to the five known cytological loci for GOE elements. Furthermore, it is unlikely that the simple structure of the GOE element could code for the different functions of these sex loci. GOE elements, because they are dispersed on all chromosomes, could be involved in measuring the X:A chromosomal balance. For example, the autosomal copies of the GOE element could interact in some way with the X chromosomal copies to assess the X:A ratio. However, this mechanism can theoretically be applied to any dispersed repeated sequence. To argue that GOE elements are more likely to be involved because they are relatively sex-chromosome sepcific in mammals and birds, fails to appreciate that the two types of sex determination are not overtly homologous.

It is possible that GOE sequences arose independently in several lineages, in which case one need not expect them to have the same function(s) in vertebrates and in insects.

Table 7.1.

Approximate cytological locations of sex loci in *Drosophila melanogaster*.

Locus name	Chromosome	Map posn.* (cM)	Cytological location** (division)
Sex lethal (Sxl)	1	19.2	6-7
daughterless (da)	2	39.3	35-36
hermaphrodite (her)	2	52.9	40
intersex (isx)	2	60.5	42-43
transformer (tra)	3	45.0	79-80
transformer-2 (tra-2)	2	70.0	46-47
doublesex (dsx)	3	48.1	81-82
maleless (mle)	2	55.2	40
male specific lethal (msl-1)	2	53.3	40
(msl-2)	2	9.0	25-26
GOE elements			11E, 19F-20AB, 38B 52F, 95A.

* Recombination map positions are obtained from Baker and Belote (1983).
** Approximate locations were determined with reference to the recombination and cytological maps published in Lindsley and Grell, (1968).

The overall conclusion from this discussion on sex determination is that GOE elements have no strong claims to being primarily involved in sex determination. Recent evidence is even more emphatic on this point, for it has been found that the cattle genome has no sequences that are detectably homologous to *Drosophila* GOE elements (Reed and Miklos, in manuscript). Obviously, GOE elements could have no significant role in mammalian sex determination if they are absent from this species. Indeed, it would be very difficult to argue that GOE elements have *any* universal or fundamental function in the light of this evidence. Therefore, there is little reason to suppose that GOE elements have an important function in preference to other dispersed and repeated sequences. Does this mean that the study of GOE elements in the *Drosophila* genome is no longer a good model system for the study of repeated sequences in general? The initial advantage that they are present in low copy number still stands. Furthermore, GOE elements probably belong to a class of dispersed and repeated simple sequences for which there is a growing body of knowledge. Thus GOE elements need not be studied in isolation and the results may be applied to a more general, and possibly important, class of repeated sequenes.

7.4 GOE elements and other dispersed, simple sequences

There are a number of short tandemly repeated simple DNA sequences that have been detected in the proximity of various genes. Some seem to be limited to single species while others are more widely represented.

The linear chromosomes of a variety of lower eukaryotes are terminated by simple, tandemly repeated sequences. These are generally of the form, C_mA_n, where *m* and *n* are positive integers (e.g. Klobutcher et al., 1981 and Blackburn and Challoner, 1984). For example, the chromosomes of four hypotrichous ciliate species are terminated by tandemly repeated CCCCAAAA sequences. They are thought to function by protecting the ends of chromosomes from degradation (Blackburn, 1984). None of the GOE elements that have been localised are situated at the telomeres of *Drosophila* chromosomes, and so they are unlikely to function in the same way as the simple terminal repeats.

Simple tandemly repeated sequences lie 5' to the expressed variable surface antigen (VSG) genes of the flagellate, *Trypanosoma brucei*. The genome contains up to 1000 VSG genes but transcription of any of these requires duplication and transposition of one copy into an 'expression site'. Thus only one gene at a time is expressed. The 5' end of the transposed segment has 1.5kb of DNA upstream of the

coding region, and this includes five tandemly arranged 70bp repeats. These are themselves made up of smaller, di- or trinucleotides (Borst, 1983). GOE elements are unlikely to be involved in switching the expression of a class of genes, because they are not flanked by homologous sequences.

B-cells proliferate and differentiate in response to an antigen. Differentiation includes altering the class of immunoglobulin molecule that is produced. This change involves the somatic recombination between homologous switch (S) regions that lie 5' to the heavy chain genes. The S region is composed of tandem repetitions of 10 to 25 copies of the pentanucleotides GAGGA and GAGGT, together with 17 interspersed copies of a different 30bp sequence. On S-S recombination, in one case, it is the 30bp repeats that are found to lie next to each other. It is not known what role the pentanucleotides play in this process (Davis et al., 1980). The switch region will be dispersed along with the immunoglobin genes. Sequences homologous to S regions have been found in *Drosophila* and other eukaryotes (Sakoyama et al., 1982). The possibility exists that these sequences are also involved in some developmentally associated recombination. If they are analogous in function to the mice sequences, they should also lie next to homologous genes. If GOE elements were to have an analogous function to these 'switch' regions, one would expect their flanking sequences to cross-hybridise, which, as discussed in Chapter 3, is not the

case.

Simple sequences have also been implicated in causing restriction fragment length polymorphisms of the human insulin gene (Bell et al., 1982). Upstream of the promoter region from this gene lies a block of tandemly repeated 14bp sequences. The number of these sequences, and therefore of the length of the relevant restriction fragment, may vary in different alleles of the insulin gene. For example, the shortest and longest RFLPs studied contain 26 and 209 14bp units, respectively. This variable block was not detectable elsewhere in the human genome and is absent from the corresponding region of the rat insulin gene.

A possibly analogous region lies 5' to the promoter site of the human myoglobin gene (Weller et al., 1984). A 650bp region is composed predominantly of GATG tetranucleotides. There are also some single base variants of this unit, including GATA. It is not known whether this region could be associated with any restriction fragment length polymorphisms. However, like the insulin gene 5' region, mentioned above, it is not detectable elsewhere in the human genome and is absent from the corresponding region of the myoglobin gene of the seal.

Simple tandemly repeated sequences can be found in the vicinity of genes in a wide range of organisms, and may be

associated with structural rearrangements of DNA. Whether all these sequences are found in other species has not, as far as is known, been studied, nor is it known how general they are within a genome.

A number of non-satellite simple sequences have been isolated from euchromatic DNA of *Drosophila virilis* and are present in other genomes (Tautz and Renz, 1984a,b). Simple repeats, such as poly(TC), poly(GT), poly(TA) and poly(CAA) were found in a number of *D. virilis* genomic clones. These clones probably contain euchromatic sequences because they were isolated from micro-libraries constructed with DNA extracted from the tips of the X and 3rd chromosomes.

Synthetic polymers of the above sequences hybridised to genomic DNA and to RNAs of a variety of eukaryotes, but not to prokaryotes. It is not known what function these simple sequences or their transcripts may have. They may be part of the coding region of genes or of their leader sequences. For example, a poly(GT) sequence lies next to one of the human actin genes and would be transcribed along with it (Hamada et al., 1982a,b).

The GOE element (poly(GATA)) probably belongs to this class of sequences. All the sequence types need not have the same function, though. For example, poly(GT) and poly(GC) sequences will form a Z-DNA structure (Wang et al., 1979). Z-DNA forms a left-handed helix and forms only one groove, whereas the B-DNA helix has both a major and a minor groove. Z-DNA sequences are thought to act by inhibiting transcription

(Rich, 1982). For example, when a sequence of DNA that is next to a promoter site and that is capable of forming Z-DNA, is removed, transcription of the downstream region is enhanced. Sequences composed of alternating purines and pyrimidines are the best candidates for forming Z-DNA structures. Poly(GATA) sequences, being three purines followed by a pyrimidine are therefore unlikely to form Z-DNA (A. Rich, pers. comm.) GOE elements may have a DNA configuration that is distinct from B-DNA (indeed, since B-DNA is an average structure, any regularly repeating sequence is likely to have a distinct structure). Whether this would be detectable crystallographically and whether it would have a functional role, are questions yet to be answered.

If poly(GATA) sequences do belong to the same class of dispersed simple sequences as do the poly(dinucleotide) sequences, one might expect to find other poly(trinucleotide) or poly(tetranucleotide) sequences dispersed in genomes. The poly(GATG) sequences, situated 5' to the human myoglobin gene have already been mentioned. A short poly(GGAA) sequence was also detected amongst a set of 300bp repeated sequences cloned from the human genome (Deininger et al., 1982). Other similar sequences, described as 'satellite-like', are situated 5' to the avian ovotransferrin and ovalbumin genes. These are of the form poly(GGAAA), poly(GGAGA) and poly(GGGAA) (Maroteaux et al., 1983). A poly(GAAA) sequence lies 5' to a human beta-actin processed gene (Moos and Gallwitz, 1983). Like the

poly(GATA) sequences, they also contain variants of the canonical sequence, whether substitutions, deletions or insertions but they have not, as yet, been analysed in the same way as has been done here. One interesting aspect of these sequences is that most of the purines are placed on one strand and most of the pyrimidines on the other. This may confer some special conformation on the DNA molecule that contains such a sequence.

As far as is known, whether all possible poly(tri- or tetra-nucleotide) sequences are represented in other eukaryote genomes has not been tested.

7.5 Possible origins and functions of dispersed and simple sequences

If dispersed simple sequences are present in all eukaryote genomes the questions to ask are - how did they arise and what are their functions, if any?

Smith (1976) demonstrated by computer simulation that a series of short unequal crossovers can generate tandemly repeated sequences in any unique sequence. Short, simple repeats would be the first to arise. This same mechanism could therefore account for the dispersed simple sequences discussed here. Satellite sequences are thought to have arisen in this way. Longer periodicities are generated when

variants spread throughout a satellite family. Why the dispersed simple sequences are not as abundant as satellite sequences could be explained by saying that they are situated in euchromatic regions which could not tolerate such excess DNA. The conclusion from this is that no function need be supposed to explain the origin of these sequences, though some may later have acquired a role.

Ohno and Epplen (1983) have argued that GOE elements could represent the remnants of sequences that were the ancestors of present day protein coding genes. A randomly generated sequence would contain a stop codon every 20 codons. Even if only one of the 64 possible triplets were initially given a terminating role, unique sequences could not, on average, generate coding regions of more than sixty codons. Simple, repeated oligonucleotides do not present this problem and could potentially generate coding sequences of any length. A poly(TATC) sequence (which is equivalent to a GOE element) could be such a primitive coding sequence. The accumulation of point mutations would allow for a gradual development of more varied polypeptides than initially coded for. Not all the *Drosophila* GOE elements look like they are translated *in vivo*, so one would have to suppose that they had lost their original coding capacities.

A contrasting view is held by Shepherd (1981). One requirement for primitive nucleotide sequences is that they code for polypeptides without the need for a start signal.

Such a code requires a specific arrangement of purines (R) and pyrimidines (Y). Shepherd (1982) suggests that the primeval codon was RNY (where N is either a purine or a pyrimidine). Any poly(RNY) sequence can provide contiguous RNY codons in one reading frame only. A poly(GATA) sequence (= RRY RRR YRR RYR) cannot be equivalent to such a primeval sequence.

7.6 Strategies for determining whether GOE elements have function

If GOE elements can be regarded as belonging to a class of dispersed, simple sequences, they can be used as a model system for a study of this class.

Most of the *Drosophila melanogaster* dispersed GOE elements and their surrounding DNAs have been cloned and identified. These can be examined in several ways to determine if they have any function. An obvious starting point is to survey the sequences for transcriptional activity. Though GOE elements do not appear to be transcribed in *Drosophila*, their flanking sequences may be. If the flanking regions are transcribed in a tissue- or stage-specific manner, this would suggest that GOE elements are control sequences. It is possible to reintegrate functioning cloned genes into the *Drosophila* genome via a P-element vector (e.g. Rubin and Spradling, 1982; Scholnick et al., 1983). The same procedure could be done for GOE elements and their

flanking regions. If a reintegrated flanking sequence without its accompanying GOE element fails to show any tissue- or stage-specificity, while the intact sequence does, this would be excellent proof that the GOE element is a control sequence.

As mentioned before, the conformation of DNA molecules containing GOE elements may be important and purified DNA can be obtained from the various plasmid subclones described in this thesis for crystallographic studies.

Of course, all of these manipulations can be carried out with any class of sequences, but the advantage here is that one need only deal with a few copies and still obtain a full description of the family within the genome. Furthermore, comparative and parallel investigations can be carried out with GOE elements from other species, such as *D. virilis* or *D. pseudoobscura*. (This cannot be done for many middle repeated sequence families, as they are limited to a few closely related species only). It would be very interesting to see how well individual GOE elements, such as GOE6, are conserved in the other species. Each member would need to be sequenced before one could say that no GOEs were equivalent in either species. Again, the advantage is that GOE is in low copy number in the *Drosophila* species.

Possibly other simple dispersed sequences are in similar copy number in the *Drosophila* genome, and similar experiments to those suggested here are as feasible. Having provided a strong data base for GOE though, this is obviously the best place to start a study of dispersed simple sequences in general.

Figure 7.1

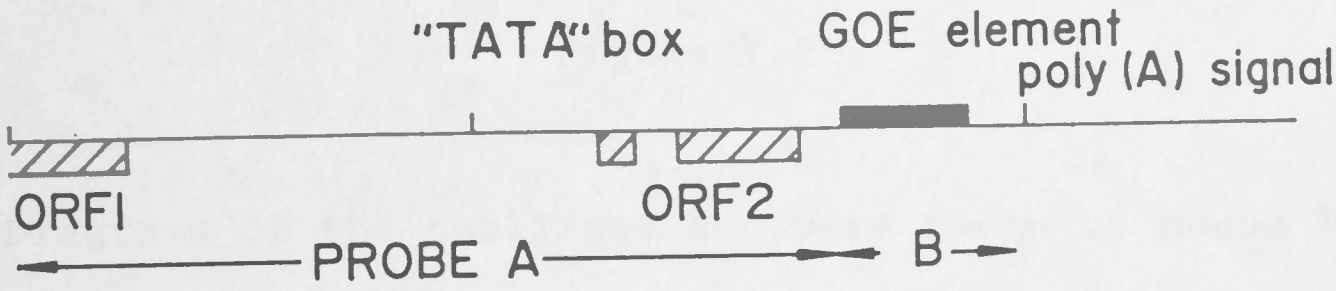
Relative positions of the GOE elements (fully shaded boxes) and of the putative open reading frames, ORFs (half-shaded boxes) in the non-*Drosophila* clones that are discussed in section 7.1.3.

A. Snake genomic clone *pErs5* from Epplen et al. (1982).

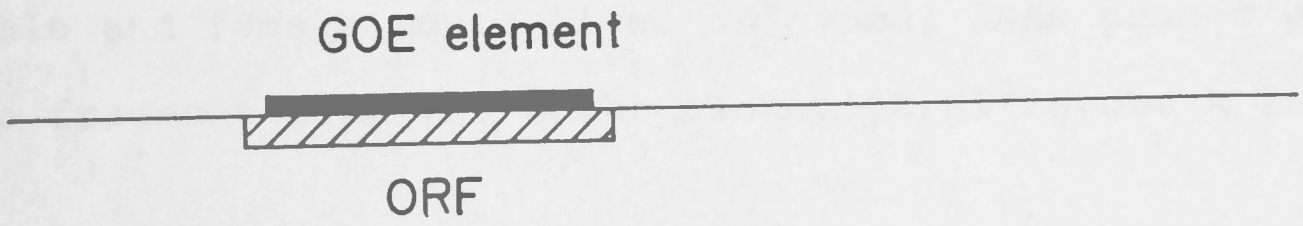
B. Mouse cDNA clone *pmc14* from Epplen et al. (1983b).

C. Mouse genomic clone *M3.1* from Singh et al. (1983).

A. Snake W-chromosome specific clone (pErs5)



B. Mouse c DNA clone (pmc 14)



C. Mouse genomic clone (M3.1)

500bp

Figure 7.2

Diagrams of the published Northern blots of mouse RNAs which have been probed with sequences containing or flanking GOE elements.

A. Male and female mouse liver *polysomal* RNAs probed with the unique fragment from the snake clone, pErs5 (probe A in figure 7.1).

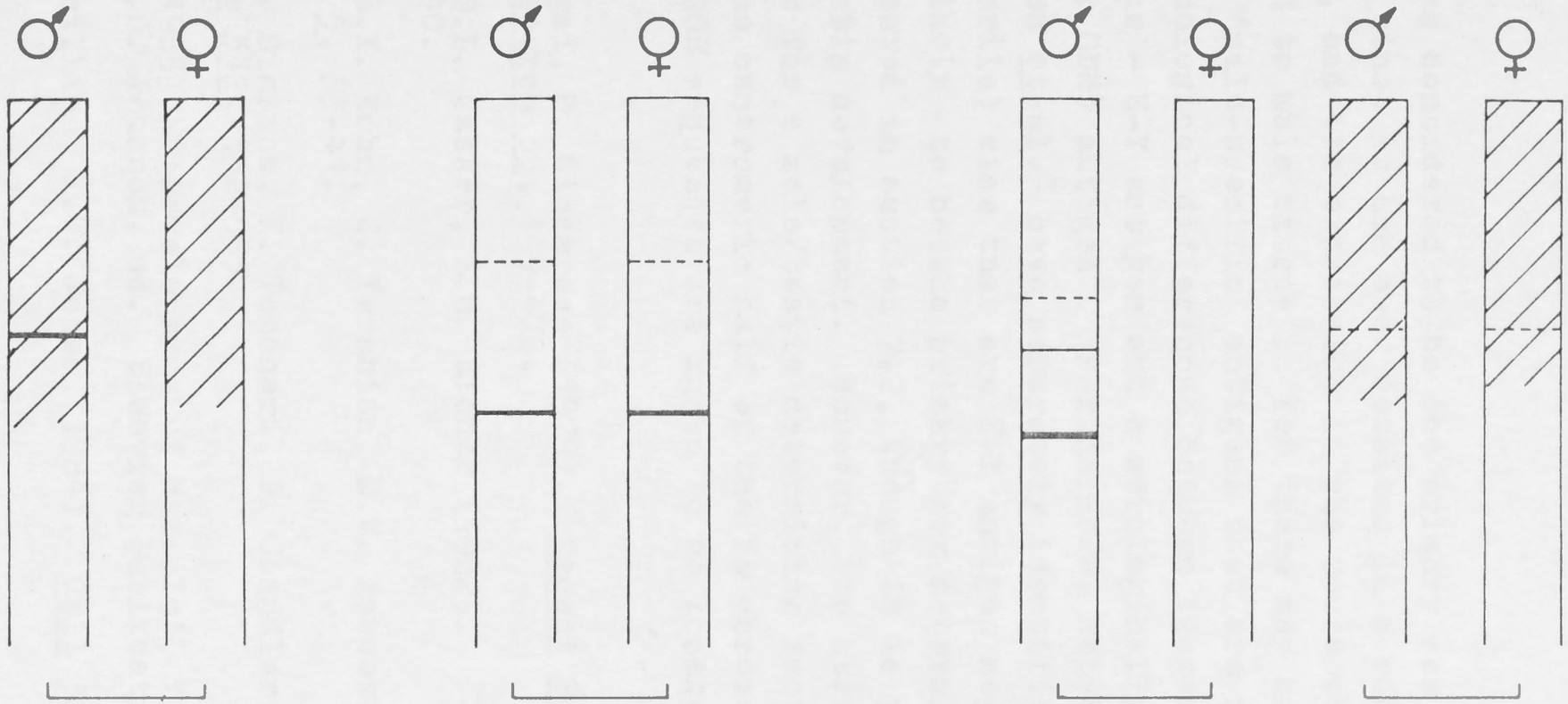
B. Male and female mouse liver *total cellular* RNAs probed with the GOE element from the snake clone, pErs5 (probe B in figure 7.1).

C. Male and female mouse liver and brain *poly(A)*$^+$ RNAs probed with the *Drosophila* GOE4 element.

A. Snake unique sequence (probe A from pErs5)

B. Snake GOE element (probe B from pErs 5)

C. Drosophila GOE4element



Polysomal RNA

Total cellular RNA

Poly(A)$^+$ RNA

ADDENDUM

H-Y antigen and sex determination

H-Y antigen was considered to be the primary sex determining factor since it had been detected in a range of vertebrate species, and its expression in the mouse was apparently confined to male tissue[1]. Yet there may be at least two separate 'male-specific' antigens that are responsible for the immunological differences between isogeneic male and female mice - H-Y antigen and a serologically determined male (or SDM) antigen[2]. Furthermore, Melvold _et al._[3] and McClaren _et al._[4] have separately identified mutant male (if sterile) mice that are H-Y antigen negative. H-Y antigen is unlikely to be the primary sex determining factor as was suggested in section 7.2, though it is probably associated with testis development. However, the structural or regulatory locus for a male/testis determining factor in mice must lie on the centromeric half of the Yp chromosome arm[5] where Bkm or GOE sequences are known to be located[6].

1. S. Ohno, Y. Nagai, S. Ciccarese (1979). _Recent Progress in Hormonal Research_ 35, 449-476.

2. W.K. Silvers, D.L. Gasser, E.M. Eicher (1982). _Cell_ 28, 439-440.

3. R.W. Melvold, H.I. Kohn, G. Yeranian, D.W. Fawcett (1977). _Immunogenetics_ 5, 33-41.

4. A. McClaren, E. Simpson, K. Tomonari, P. Chandler, H. Hogg (1984). _Nature_ 312, 552-555.

5. A.D. Stewart (1983). 'Development of Mammals', Vol. 5, Chapter 10. M.H. Johnson, ed. Elsevier Publications B.V.

6. L. Singh, C. Phillips, K.W. Jones (1984). _Cell_ 36, 111-120.

REFERENCES CITED

Alonso, A., B. Kuehn, J. Fischer (1983). An unusual accumulation of repetitive sequences in the rat genome. Gene 26, 303-306.

Ananiev, E.V., V.A. Gvozdev, Yu.V. Ilyin, N.A. Tchurikov, G.P. Georgiev (1979). Reiterated genes with varying location in intercalary heterochromatin regions of *Drosophila melanogaster* polytene chromosomes. Chromosoma 70, 1-17.

Baker, B.S. and J.M. Belote (1983). Sex determination and dosage compensation in *Drosophila melanogaster*. Ann. Rev. of Genetics 17, 345-393.

Baldari, C.T. and F. Amaldi (1977). DNA reassociation kinetics in relation to genome size in four Amphibian species. Chromosoma 59, 13-22.

Bell, G.I., M.J. Selby, W.J. Rutter (1982). The highly polymorphic region near the human insulin gene is composed of simple tandemly repeated sequences. Nature 295, 31-35.

Benton, W.D. and R.W. Davis (1977). Screening of lambda gt10 recombinant clones by hybridisation to single plaques *in situ*. Science 196, 180-182.

Blackburn, E.H. (1984). Telomeres: do the ends justify the means? Cell 37, 7-8.

Blackburn, E.H. and P.B. Challoner (1984). Identification of a telomeric DNA sequence in *Trypanosoma brucei*. Cell 36, 447-457.

Bolivar, F., R.L. Rodriguez, P.J. Greene, M.C. Betlach, H.L. Heynecker, H.W. Boyer, J.H. Crosa, S. Falkow (1977). Construction and characterisation of new cloning vehicles. II. A multipurpose cloning system. Gene 2, 95-113.

Borchsenius, S.N, N.A. Belozerskaya, N.A. Merkulova, V.G. Wolfson, V.I. Vorob'ev (1978). Genome structure of *Tetrahymena pyriformiis*. Chromosoma 69, 275-289.

Borst, P. (1983). Antigenic variation in

trypanosomes. In: 'Mobile Genetic Elements', (J.A. Shapiro, ed.), pp. 619-656. Academic Press, New York.

Britten, R.J. and D.E. Kohne (1968). Repeated sequences in DNA. Science 161, 249-256.

Brown, A.H.D. and M.T. Clegg (1983). Analysis of variation of related DNA sequences. In: 'Statistical Analysis of DNA Sequence Data', (B.S. Weir, ed.), Chapter 5. Marcel Dekker, New York.

Brown, S.D. and G. Dover (1981). Organisation and evolutionary progress of a dispersed repetitive family of sequences in widely separated rodent genomes. J. Mol. Biol. 150, 441-466.

Brown, S.D. and M. Piechaczyk (1983). Insertion sequences and tandem repetitions as sources of variation in a dispersed repeat family. J. Mol. Biol. 165, 249-256.

Brutlag, D.L. (1980). Molecular arrangement and evolution of heterochromatic DNA. Ann. Rev. of Genetics 14, 121-144.

Brutlag, D.L. and W.J. Peacock (1979). DNA sequences of the 1.672g/cm^3 satellite of *Drosophila melanogaster*. J. Mol. Biol. 135, 565-580.

Bukhari, A. and L. Ambrosio (1978). The invertible segment of bacteriophage Mu DNA determines the adsorption properties of Mu particles. Nature 271, 575-577.

Calabretta, B., D.L. Robberson, H.A. Barrera-Saldena, T.P Lambron, G.F. Saunders (1982). Genomic instability in a region of human DNA enriched in Alu repeat sequences. Nature 296, 219-225.

Calos, M.P. and J.H. Miller (1980). Transposable elements. Cell 24, 753-763.

Carroll, D., J.E. Garrett, B.S. Lan (1984). Isolated clusters of paired tandemly repeated sequences in the *Xenopus laevis* genome. Mol. and Cell. Biology 4, 254-259.

Christie, N.T. and D.M. Skinner (1979). Interspersion of highly repetitive DNA with single copy DNA in the genome of the red crab, *Geryon quinquedens*. Nucleic Acids Res. 6, 781-796.

Chung, S., C. Zuker, H. Lodish (1983). A repetitive and apparently transposable DNA sequence in *Dictyostelium discoideum* associated with developmentally regulated mRNAs. Nucleic Acids Res. 11, 4835-4852.

Clarkson, S.G., M.L. Birnstiel, I.F. Purdom (1973). Clustering of transfer RNA genes of *Xenopus laevis*. J. Mol. Biol. 79, 411-429.

Coen, E.S., J.M. Thoday, G. Dover (1982). Rate of turnover of structural variants in the rDNA gene family of *Drosophila melanogaster*. Nature 295, 564-568.

Cooke, H.J., J. Schmidtke, J.R. Gosden (1982). Characterisation of a human Y chromosome repeated sequence and related sequences in higher primates. Chromosoma 87, 491-502.

Costantini, F.D., R.H. Scheller, R.J. Britten, E.H. Davidson (1978). Repetitive sequence transcripts in the mature sea urchin oocyte. Cell 15, 173-187.

Crain, W.R., E.H. Davidson, R.J. Britten (1976). Contrasting patterns of DNA sequence arrangement in *Apis mellifera* (Honeybee) and *Musca domestica* (Housefly). Chromosoma 59, 1-12.

CRC (Chemical Rubber Company) Handbook of Biochemistry; selected data for molecular biolgy, A.H. Sober, ed. 1968.

Davidson, E.H. and R.J. Britten (1973). Organisation, transcription and regulation in the animal genome. Quart. Rev. Biol. 48, 565-613.

Davidson, E.H. and R.J. Britten (1979). Regulation of gene expression: possible role of repetitive sequences. Science 204, 1052-1059.

Davidson, E.H. and J.W. Posakony (1982). Repetitive sequence transcripts in development. Nature 284, 633-635.

Davidson, E.H., B.R. Hough, C.S. Amenson, R.J. Britten (1973a). General interspersion of repetitive and non-repetitive sequence elements in the DNA of *Xenopus*. JMB 77, 1-23.

Davidson, E.H., B.R. Hough, W.H. Klein, R.J. Britten (1975). Structural genes adjacent to interspersed repetitive DNA sequences. Cell 4, 217-238.

Davidson, E.H., H.T. Jacobs, R.J. Britten (1983). Very short repeats and coordinate induction of genes. Nature 301 468-470.

Davis, M.M., S.K. Kim, L.E. Hood (1980). DNA sequences mediating class switching in alpha-immunoglobulins. Science 309, 1360-1365.

Deininger, P.L., D.J. Jolly, C.M. Rubin, T. Friedmann, C.W. Schmid (1981). Base sequence studies of 300 nucleotide renatured repeated human DNA clones. JMB 151, 17-33.

Diaz, M.O., G. Barsacch-Pilone, K.A. Mahon, J.G. Gall (1981). Transcripts from both strands of a satellite DNA occur on lampbrush chromosome loops of the newt, *Notophthalmus*. Cell 24, 649-659.

DiGiovanni, L., S.R. Haynes, R. Misra, W.R. Jelinek (1983). Kpn I family of long-dispersed repeated DNA sequences of man: Evidence for entry into genomic DNA of DNA copies of poly(A)-terminated Kpn I RNAs. PNAS 80, 6533-6537.

DiNocera, P.P., M.E. Digan, I.B. Dawid (1983). A family of oligo-adenylate-terminated transposable sequences in *Drosophila melanogaster*. J. Mol. Biol. 168, 715-727.

Doolittle, W.F. and C. Sapienza (1980). Selfish genes, the phenotype paradigm and genome evolution. Nature 284, 601-603.

Dover, G. (1980). Ignorant DNA? Nature 285, 618-620.

Dover, G. (1982). Molecular drive: a cohesive model of species evolution. Nature 299. 111-117.

Dowsett, A.P. and M.W. Young (1982). Differing levels of dispersed repetitive DNA among closely related species of *Drosophila*. PNAS 79, 4570-4574.

Dowsett, A.P. (1983). Closely related species of *Drosophila* can contain different libraries of middle repetitive DNA sequences. Chromosoma 88, 104-108.

Eibel, H., J. Gafner, A. Stotz, P. Philippsen (1980). Characterisation of the yeast mobile element Ty1. Cold Spring Harbor Sym. Quant. Biol. 45(2), 609-617.

Emmons, S.W., L. Yesner, K.-S. Ruan, D. Katzenberg (1983). Evidence for a transposon in *Caenorhabditis elegans*. Cell 32, 55-65.

Endow, S.A. and J.G. Gall (1975). Differential replication of satellite DNA in polyploid tissues of *Drosophila virilis*. Chromosoma 50, 175-179.

Endow, S.A., M.L. Polan, J.G. Gall (1975). Satellite DNA sequences of *Drosophila melanogaster*. J. Mol. Biol. 96, 665-692.

Epplen, J.T., M. Leipoldt, W. Engel, J. Schmidtke (1978). DNA sequence organisation in avian genomes. Chromosoma 69, 307-321.

Epplen, J.T., S.Sutou, J.R. McCarrey, S Ohno (1981). Sex-determining genes and gene regulation. In: 'Fetal Endocrinology: Oregon Regional Primate Research Centre Symposia on primate reproductive biology', (M.J. Novy and J.A. Resko, eds), pp. 239-251. Academic Press, New York.

Epplen, J.T., J.R. McCarrey, S. Sutou, S. Ohno (1982). Base sequence of a cloned snake W-chromosome fragment and identification of a male-specific putative mRNA in the mouse. PNAS 79, 3798-3802.

Epplen, J.T. A. Cellini, M. Shorte, S. Ohno (1983a). On evolutionarily conserved simple repetitive DNA sequences: do 'sex-specific' satellite components serve any sequence dependent function? Differentiation 23(S), S60-S63.

Epplen, J.T., A. Cellini, S. Romero, S. Ohno (1983b) An attempt to approach the molecular mechanisms of primary sex determination: W- and Y-chromosomal conserved simple repetitive DNA sequences and their differential expression in mRNA. J. Exp. Zoology 228, 305-312.

Fanning, T.G. (1982). Characterisation of a highly repetitive family of DNA sequences in the mouse. Nucleic Acids Res. 10, 5003-5013.

Finnegan, D.J., G.M. Rubin, M.W. Young, D.S. Hogness (1977). Repeated gene families in *Drosophila melanogaster*. Cold Spring Harbor Sym. Quant. Biol. 42(2), 1053-1063.

Finnegan, D.J., B.H. Will, A.A. Bayev, A.M. Bostock, L. Brown (1982). Transposable DNA sequences in eukaryotes. In: 'Genome Evolution' (G.A. Dover and R.B. Flavell, eds.), pp. 27-40. Academic Press, London.

Flavell A.J. and D. Ish-Horowicz (1981). Extrachromosomal circular copies of the eukaryotic transposable element *copia* in cultured *Drosophila* cells. Nature 292, 591-595.

Flavell R.B., M.D. Bennett, J.B. Smith, D.B. Smith (1974). Genome size and the proportion of repeated nucleotide sequence DNA in plants. Biochem. Genetics 12, 257-269.

Frankham, R., D.A. Briscoe, R.K. Nurthen (1980). Unequal crossing over at the rRNA tandon as a source of quantitative genetic variation in *Drosophila*. Genetics 95, 727-742.

Fry, K. and D.L. Brutlag (1979). Detection and resolution of closely related sequences in satellite DNA by molecular cloning. J. Mol. Biol. 135, 581-593.

Fyrberg, E.A., K.L. Kindle, N. Davidson, A. Sodja (1980). The actin genes of *Drosophila*: a dispersed multigene family. Cell 19, 365-378.

Galau, G.A., M.E. Chamberlin, B.R. Hough, R.J. Britten, E.H. Davidson (1976). Evolution of repetitive and non-repetitive DNA. In: 'Molecular Evolution', (F.J. Ayala., ed.), pp. 200-224. Sinauer Associates, Sunderland.

Gebhard, W., T. Meitinger, J. Hoechtl, H.G. Zachau (1982). A new family of interspersed repetitive DNA sequences in the mouse genome. J. Mol. Biol. 157, 53-471.

Gebhard, W. and H.G. Zachau (1983). Organisation of the R family and other interspersed repetitive DNA sequences in the mouse genome. J. Mol. Biol. 170, 255-270.

Georgiev, G.P., Y.V. Ilyin, V.G. Chmeliauskaite, A.P. Ryskov, D.A. Kramerov, K.G. Skryabin, A.S Krayev, E.M. Lukanidin, M.S. Grigoryan (1980). Mobile dispersed genetic elements and other middle repetitive DNA sequences in the genomes of *Drosophila* and mouse: Transcriptional and biological significance. Cold Spring Harbor Sym. Quant. Biol. 45(2), 641-654.

Georgiev, G.P., D.A. Kramerov, A.P. Ryskov, K.G Skryabin, E.M. Lukanidin (1983). Dispersed repetitive sequences in eukaryotie genomes and their possible biological significance. Cold Spring Harbor Sym. Quant. Biol. 47(2), 1109-1121.

Gilroy, T.E and C.A. Thomas, Jr. (1983). The analysis of some new *Drosophila* repetitive DNA sequences isolated and cloned from two-dimensional gels. Gene 23, 41-51.

Goldberg, D.A. (1980). Isolation and partial characterisation of the *Drosophila* alcohol dehydrogenase gene. PNAS 77, 5794-5798.

Goldberg, R.B., W.R Crain, J.V. Ruderman, G.P. Moore, T.R. Barnett, R.C. Higgins, R.A Gelfand, G.A. Galau, R.J. Britten, E.H. Davidson (1975). DNA sequence organisation in the genomes of five marine invertebrates. Chromosoma 51, 225-251.

Hamada, H., M.G. Petrino, T. Kakunaga (1982a). Molecular structure and evolutionary origin of human cardiac muscle actin gene. PNAS 79, 5901-5905.

Hamada, H., M.G. Petrino, T. Kakunaga (1982b). A novel repeated element with Z-DNA-forming potential is widely found in evolutionarily diverse eukaryote genomes. PNAS 79, 6465-6469.

Haynes, S.R. and W.R. Jelinek (1981). Low molecular weight RNAs transcribed *in vitro* by RNA polymerase III from Alu-type dispersed repeats in Chinese hamster are also found *in vivo*. PNAS 78, 6130-6134.

Hershey, N.D., S.E. Conrad, A. Sodja, P.H. Yen, M. Cohen, Jr., N. Davidson, C. Ilgen, J. Carbon (1977). The sequence arrangement of *Drosophila melanogaster* 5S DNA cloned in recombinant plasmids. Cell 11, 585-598.

Hsieh, T. and D.L. Brutlag (1979). Sequence and sequence variation within the 1.688g/cm^3 satellite DNA of *Drosophila melanogaster*. J. Mol. Biol. 135, 465-481.

Hudspeth, M.E.S., W.E. Timberlake, R.B. Goldberg (1977). DNA sequence organisation in the water mold, *Achyla*. PNAS 74 4332-4336.

Ising, G. and K. Block (1980). Derivation-dependent

distribution of insertion sites for a *Drosophila* transposon. Cold Spring Harbor Symp. Quant. Biol. 45(2), 527-544.

Jagadeeswaran, P., D. Tuan, B.G Forget, S.M. Weissman (1982). A gene deletion ending at the midpoint of a repetitive DNA sequence in one form of hereditary persistence of foetal haemoglobin. Nature 296, 469-470.

Jelinek, W.R. and S.R. Haynes (1983). The mammalian Alu family of dispersed repeats. Cold Spring Harbor Sym. Quant. Biology 47(2), 1123-1130.

John, B. and G.L.G. Miklos (1979). Functional aspects of satellite DNA and heterochromatin. Int. Rev. of Cytology 58, 1-114.

Jones, K.W. (1983). Evolutionary conservation of sex specific DNA sequences. Differentiation 23(S), S56-S59.

Jones, K.W. and L. Singh (1981). Conserved repeated DNA sequences in vertebrate sex chromosomes. Human Genetics 38, 46-53.

Jones, K.W. and L. Singh (1982). Conserved sex associated repeated DNA sequences in vertebrates. In: 'Genome Evolution' (Flavell, A.J. and Dover, G., eds.), pp. 135-154. Academic Press, New York.

Kay, B.K. and I.B. Dawid (1983). The 1723 element: a long, homogeneous, highly repeated DNA unit interspersed in the genome of *Xenopus laevis*. J. Mol. Biol. 170, 583-596.

Kedes, L.H. (1979). Histone genes and histone messengers. Ann. Rev. Biochem. 48, 837-870.

Kimmel, A.R. and R.A. Firtel (1979). A family of short, interspersed repeat sequences at the 5' end of a set of *Dictyostelium* single copy mRNAs. Cell 16, 787-796.

Klein, H.L. and T.D. Petes (1981). Intrachromosomal gene conversion in yeast. Nature 289, 144-148.

Klobutcher, L.A., M.T. Swanton, P. Donini, D.M. Prescott (1981). All gene-sized DNA molecules in four species of hypotrichs have the same terminal sequence and an unusual 3' terminus. PNAS 78, 3015-3019.

Kominami, R., M. Muramatsu, K. Moriwaki (1983).

Novel repetitive sequence families showing size and frequency polymorphisms in the genomes of mice. Nature 301, 87-89.

Levis, R., M. Collins, G.M. Rubin (1982). FB elements are the common basis for the instability of the w^{DZL} and w^{C} *Drosophila* mutations. Cell 30, 551-565.

Lewin, B. (1980). Sequences of eukaryote DNA. In: 'Gene Expression', Volume 2, pp. 148-228. Wiley, New York.

Lewis, N. and J.B. Gibson (1978). Variation in amount of enzyme protein in natural populations. Biochem. Genetics 16,159-170.

Liebermann, D., B. Hoffmann-Liebermann, J. Weinthal, G. Childs, R. Maxson, A. Mauron, S.N. Cohen, L. Kedes (1983). An unusual transposon with long terminal inverted repeats in the sea urchin, *Strongylocentrotus purpuratus*. Nature 306, 342-347.

Lindsley, D.L. and E.H. Grell (1968). Genetic variations in *Drosophila melanogaster*. Carnegie Institute of Washington Publication (No. 627).

Long, E.O. and I.B. Dawid (1980). Repeated genes in eukaryotes. Ann. Rev. Biochem. 49, 727-764.

Manning, J.E., C.W. Schmid, N. Davidson (1975). Interspersion of repetitive and non-repetitive DNA sequences in the *Drosophila melanogaster* genome. Cell 4, 141-155.

Maniatis, T., R.C. Richardson, E. Lacy, J.Lauer, C. O'Connell, D. Quon, G.K. Sim, A. Efstratiadis (1978). The isolation of structural genes from libraries of eukaryotic DNA. Cell 15, 687-701.

Maroteaux, L., R. Heilig, D. Dupret, J.L. Mandel (1983). Repetitive satellite-like sequences are present within or upstream from 3 avian protein-coding genes. Nucleic Acids Res. 11, 1227-1243.

Martin, G., D. Wiernasz, P. Schedl (1983). Evolution of *Drosophila* repetitive dispersed DNA. J. Mol. Biol. 19, 203-213.

McCarrey, J.R. and U.K. Abbott (1979). Mechanisms of genetic sex determination, gonadal sex differentiation and germ-cell development in animals. Adv. in Genetics 20, 217-290.

McCarthy, B.J. and R.B. Church (1970). The specificity of molecular hybridisation reactions. Ann. Rev. Biochemistry 39, 131-150.

McGinnis, W., M.S. Levine, E. Hafen, A. Kuroiwa, W.J. Gehring (1984). A conserved DNA sequence in homeotic genes of the *Drosophila* Antennapedia and bithorax complexes. Nature 304, 428-433.

Messing, J. and J. Vieira (1982). A new pair of M13 vectors for selecting either DNA strand of double-digest restriction fragments. Gene 19, 269-276.

Messing, J., B. Gronenborn, B. Mueller-Hill, P.H. Hofschneider (1977). Filamentous coliphage M13 as a cloning vehicle: Insertion of a Hind II fragment of the *lac* regulatory region in M13 replicative form *in vitro*. PNAS 74, 3642-3646.

Messing, J., R. Crea, P.H. Seeburg (1981). A system for shotgun sequencing. Nucleic Acids Res. 9, 309-321.

Miller, J.H. (1972). 'Experiments in Molecular Genetics'. Cold Spring Harbor Laboratory, Cold Spring Harbor, New York.

Miklos, G.L.G., M.J. Healy, P. Pain, A.J. Howells, R.J. Russell (1984). Molecular and genetic studies on the euchromatin-heterochromatin transition region of the X chromosome of *D. melanogaster*. I. A cloned entry point near to the uncoordinated (*unc*) locus. Chromosoma 89, 218-227.

Moore, G.P., A.R. Moore, L.I. Grossman (1984). The frequency of matching sequences in DNA. J. Theor. Biol. 108, 111-122.

Moos, M. and D. Gallwitz (1983). Structure of two human beta-actin-related processed genes, one of which is located next to a simple repetitive sequence. EMBO Journal 2, 757-761.

Murray, N.E., W.J. Brammer, K. Murray (1977). Lambdoid phages that simplify recovery of *in vitro* recombinants. Mol. Gen. Genetics 150, 53-61.

Musti, A.M., D.A. Sobieski, B.B. Chen, F.C. Eden (1981). Repeated DNA clusters in the chicken genome contain homologous sequence elements in scrambled order. Biochemistry 20, 2899-2999.

Needleman, S.B. and C.D. Wunsch (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. 48, 443-453.

O'Hare, K. and G.M. Rubin (1983). Structures of P elements and their sites of insertion and excision in the *Drosophila melanogaster* genome. Cell 34, 25-35.

Ohno, S. (1967). Sex chromosomes and sex-linked genes. Springer, Berlin.

Ohno, S. and J.T. Epplen (1983). The primitive code and repeats of base oligomers are the primordial protein-encoding sequence. PNAS 80, 3391-3395.

Ohtsubo, H. and E. Ohtsubo (1977). Repeated DNA sequences in plasmids, phages and bacterial chromosomes. In: 'DNA Insertion Elements, Plasmids and Episomes', (A.I. Bukhari, J.A. Shapiro and S.L. Adhya, eds.). pp. 49-63. Cold Spring Harbor Lab., Cold Spring Harbor, New York.

Orgel, L.E. and F.H.C. Crick (1980). Selfish DNA: the ultimate parasite. Nature 284, 604-607.

Ottolenghi, S. and B. Giglioni (1982). The deletion of a type of delta-beta-thalassaemia begins in an inverted Alu I repeat. Nature 300, 770-771.

Patterson, J.T. and W.S. Stone (1952). Evolution in the genus Drosophila. Macmillan, New York.

Peoples, O.P. and N. Hardman (1983). An abundant famliy of methylated repetitive sequence dominates the genome of *Physarum polycephalum*. Nucleic Acids Res. 11, 7777-7788.

Potter, S.S. (1982). DNA sequence of a foldback transposable element in *Drosophila*. Nature 297, 201-204.

Rich, A. (1982). Right-handed and left-handed DNA: Conformational information in genetic material. Cold Spring Harbor Sym. Quant.Biol. 47(1), 1-12.

Rogers, J. (1983). A straight LINE story. Nature 306, 113-114.

Rubin, G.M. (1983). Dispersed repetitive DNAs in *Drosophila*. In: 'Mobile Genetic Elements', (ed. J.A. Shapiro), pp. 329-361. Academic Press, New York.

Rubin, G.M. and A.C. Spradling (1982). Genetic transformation of *Drosophila* with transposable element vectors. Science 218, 348-353.

Rubin, G.M., W.J. Brorein, Jr., P. Dunsmuir, A.J. Flavell, R. Levis, E. Strobel, J.L. Toole, E. Young (1980). *Copia*-like transposabe elements in the *Drosophila* genome. Cold Spring Harbor Sym. Quant. Biol. 45(2), 629-640.

Sakoyama, Y., Y. Yaoita, T. Honjo (1982). Immunoglobulin switch-region-like sequences in *Drosophila melanogaster*. Nucleic Acids Res. 10, 4203-4214.

Sampsell, B. (1977). Isolation and genetic characterisation of alcohol dehydrogenase thermostability variants in natural populations of *Drosophila melanogaster*. Biochem. Genetics 15, 971-988.

Sanger, F., S. Nicklen, A.R. Coulson (1977). DNA sequencing with chain-terminating inhibitors. PNAS 74, 5463-5467.

Sapienza, C. and W.F. Doolittle (1982). Unusual physical organisation of the *Halobacterium* genome. Nature 295, 384-389.

Schalet, A. (1969). Exchanges at the bobbed locus of *Drosophila melanogaster*. Genetics, 63, 133-153.

Schmid, C.W. and W.R. Jelinek (1982). The Alu family of dispersed repetitive sequences. Science 216, 1065-1070.

Schmidtke, J. and J.T. Epplen (1980). Sequence organisation of animal nuclear DNA. Human Genetics 55, 1-18.

Scholnick, S.B., B.A. Morgan, J. Hirsch (1983). The cloned dopa-decarboxylase gene is developmentally regulated when reintegrated into the *Drosophila* genome. Cell 34, 37-45.

Shepherd, J.C.W. (1981). Method to determine the

reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. PNAS 78, 1596-1600.

Shepherd, J.C.W. (1982). From primeval message to present-day gene. Cold Spring Harbor Sym. Quant. Biol. 47(2), 1099-1108.

Shiba, T. and K. Saigo (1983). Retrovirus-like particles containing RNA homologous to the transposable element *copia* in *Drosophila melanogaster*. Nature 302, 119-124.

Silvermann, M., J. Zieg, M. Hilmen, M. Simon (1979). Phase variation in *Salmonella*: Genetic analysis of a recombinational switch. PNAS 76, 391-395.

Singer, M.F. (1982a). Highly repeated sequences in mammalian genomes. Int. Rev. of Cytology 76, 67-112.

Singer, M.F. (1982b). SINEs and LINEs: Highly repeated short and long interspersed sequences in mammalian genomes. Cell 28, 433-434.

Singer, M.F., R.E. Thayer, G. Grimaldi, M.I. Lerman, T.G. Fanning (1983). Homology between the KpnI primate and BamHI (MIF-1) rodent families of long interspersed repeated sequences. Nucleic Acids Res. 11, 5739-5745.

Singh L. and K.W. Jones (1982). Sex reversal in the mouse (Mus musculus) is caused by a recurrent nonreciprocal crossover involving the X and an aberrant Y chromosome. Cell 28, 205-216.

Singh, L., I.F. Purdom, K.W. Jones (1976). Satellite DNA and evolution of sex chromosomes. Chromosoma 59, 43-62.

Singh, L., I.F. Purdom, K.W. Jones (1980a). Sex-chromosome-associated satellite DNA: Evolution and conservation. Chromosoma 79, 137-157.

Singh, L., I.F. Purdom, K.W. Jones (1980b). Conserved sex-chromosome-associated nucleotide sequences in eukaryotes. Cold Spring Harbor Sym. Quant. Biol. 45(2), 805-813.

Singh, L., C. Phillips, K.W. Jones (1984). The conserved nucleotide sequences of Bkm, which define *Sxr* in the mouse, are transcribed. Cell

36, 111-120.

Smith, G.P. (1976). Evolution of repeated DNA sequences by unequal crossovers. Science 191, 528-535.

Soberon, X., L. Covarrubias, F. Bolivar (1980). Construction and characterisation of new cloning vehicles. IV. Deletion derivatives of pBR322 and pBR325. Gene, 9, 287-305.

Spierer, P., A. Spierer, W. Bender, D.S. Hogness (1983). Molecular mapping of genetic and chromomeric units in *Drosophila melanogaster*. J. Mol. Biol. 168, 35-50.

Spohr, G., W. Reith, I. Sures (1981). Organisation and sequence analysis of repetitive DNA elements from *Xenopus laevis*. J. Mol. Biol. 151, 573-592.

Streeck, R.E. (1982). A multicopy insertion sequence in the bovine genome with structural homology to the long terminal repeats of retroviruses. Nature 298, 767-769.

Sueoka, N. and T.Y. Cheng (1962). Natural occurrence of a deoxyribonucleic acid resembling the deoxy-adenylate-deoxythymidylate polymer. PNAS 48, 1851-1856.

Sun, L., K.E. Paulson, C.W. Schmid, L. Kadyk, L. Leinwand (1984). Non-Alu family interspersed repeats in human DNA and their transcriptional activity. Nucleic Acids Res. 12, 2669-2690.

Sutton, W.D., W.L. Gerlach, D. Schwartz, W.J. Peacock (1984). Molecular analysis of Ds controlling element mutations at the *Adh 1* locus of maize. Science 223, 1265-1268.

Tautz, D. and M Renz (1984a). Simple DNA sequences of *Drosophila virilis* isolated by screening with RNA. J. Mol. Biol. 172, 229-235.

Tautz, D. and M. Renz (1984b). Simple sequences are ubiquitous repetitive components of eukaryote genomes. Nucleic Acids Res. 12, 4127-4138.

Tchurikov, N.A., A.K. Naumova, E.S. Zelentsova, G.P. Georgiev (1982). A cloned unique gene of *Drosophila melanogaster* contains a repetitive 3' exon whose sequence is present at the 3' ends of many different mRNAs. Cell 28, 365-373.

Timberlake, W.E. (1978). Low repetitive DNA content in *Aspergillus nidulans*. Science 202, 973-975.

Truett, M.A., R.S. Jones, S.S. Potter (1981). Unusual structure of the FB family of transposable elements in *Drosophila*. Cell 24, 753-763.

Tone, M., N. Nakano, E. Takao, S. Narisawa, S. Mizuno (1982). Demonstration of W chromosome-specific repetitive DNA sequences in the domestic fowl, *Gallus g. domesticus*. Chromosoma 86, 229-235.

Tone, M., Y. Sakaki, T. Hashiguchi, S. Mizuno (1984). Genus specificity and extensive methylation of the W chromosome-specific repetitive DNA sequences from the domestic fowl, *Gallus g. domesticus*. Chromosoma 89, 551-569.

Ullu, E. (1982). The human Alu family of repeated DNA sequences. Trends in Biochemical Sciences 7, 216-219.

Varley, J.M., H.C. Macgregor, H.P. Erba (1980). Satellite DNA is transcribed on lampbrush chromosomes. Nature 283, 686-688.

Varmus, H.E. (1983). Retroviruses. In: 'Mobile Genetic Elements', (J.A. Shapiro, ed.), pp. 411-503. Academic Press, New York.

Wachtel, S.S., G.C. Koo, E.A. Boyse (1975). Evolutionary conservation of H-Y ('male') antigen. Nature 254, 270-272.

Wang, A.H.-J., G.J. Quigley, F.J. Kolpak, J.L. Crawford, J.H. van Boom, G. van der Marel, A. Rich (1979). Molecular structure of a left-handed double helical DNA fragment at atomic resolution. Nature 282, 680-686.

Weller, P., A.J. Jeffreys, V. Wilson, A. Blanchetot (1984). Organisation of the human myoglobin gene. EMBO Journal 3, 439-446.

Wensink, P.C., S. Tabata, C. Pachl (1979). The clustered and scrambled arrangement of moderately repetitive elements in *Drosophila* DNA. Cell 18, 1231-1246.

Whitfeld, P.L., P.H. Seeburg, J. Shine (1982). The human proopiomelanocortin gene: organisation, sequence and interspersion with repetitive DNA. DNA 1, 133-143.

Wimber, D.E. and D.M. Steffensen (1970). Localisation of 5S RNA genes on *Drosophila* chromosomes by RNA-DNA hybridisation. Science 170, 639-645.

Wirth, T. K., Gloeggler, T. Baumruker, M Schmidt, I. Horak (1983). Family of middle repetitive DNA sequences in the mouse genome with structural features of solitary retroviral long terminal repeats. PNAS 80, 3327-3330.

Yen, P.H. and N. Davidson (1980). The gross anatomy of a tRNA cluster at region 42A of the *Drosophila melanogaster* genome. Cell 22, 137-148.

Young, M.W. (1979). Middle repetitive DNA: a fluid component of the *Drosophila* genome. PNAS 76, 6274-6278.

Young, M.W. and H.E. Schwartz (1980). Nomadic gene families in *Drosophila*. Cold Spring Harbor Sym. Quant. Biol. 45(2), 629-640.

Zhimulev, I.F., V.F. Shemeshin, V.A. Kulichov, E.S. Belyaeva (1982). Intercalary heterochromatin in *Drosophila*. 1. Localisation and general characteristics. Chromosoma 87, 197-228.

Zuker, C. and H.F. Lodish (1981). Repetitive DNA sequences co-transcribed with developmentally regulated *Dictyostelium discoideum* mRNAs. PNAS 78, 5386-5390.